ESTIMATION OF LOCATION AND SCALE PARAMETERS AND THEIR JOINT DENSITY IN A RANDOM EFFECTS MODEL

A Thesis

by

SHYAMALENDU SINHA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jeffrey D. Hart
Committee Members,	Valen E. Johnson
	Mohsen Pourahmadi
	Sue Geller
Head of Department,	Valen E. Johnson

December 2018

Major Subject: Statistics

ABSTRACT

This thesis provides a framework for estimating the location-scale parameters in random effects models. A secondary goal, which is necessary to efficiently achieve the main goal, is to estimate the joint density of the location-scale parameters.

The main setting considered here is having a large number of small data sets whose locations and scales vary randomly but have a common joint distribution. The goal is to estimate the location-scale parameters and their joint density assuming the scaled error density is standard normal. This thesis relaxes the assumption that location and scale are independent and introduces a Bayesian semi-parametric approach based on a mixture of normal-inverse gamma densities. Also, this thesis further relaxes the assumption that the scaled error density is standard normal, instead allowing any known scaled error density. The joint density of location and scale is estimated by a bivariate histogram. Estimation algorithms are proposed and their usefulness is illustrated with both simulated and real data.

DEDICATION

To my mother, my father, my grandfathers, and my grandmothers.

ACKNOWLEDGMENTS

The long journey through the graduate school would not be possible without kind support of friends and families.

First, I am greatly thankful to my advisor, Dr. Jeffrey D. Hart, for his patience and valuable guidance. I have been fortunate to have him as my advisor. I would like to express special gratitude to Dr. Valen E. Johnson, Dr. Mohsen Pourahmadi and Dr. Sue Geller for their time and suggestion while serving as members of my committee.

I would like to thank Dr. Michael Longnecker and Department of Statistics, Texas A&M University for supporting me as a Teaching Assistant for the entire duration of the graduate school.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis (or) dissertation committee consisting of Professor Jeffrey D. Hart, Valen E. Johnson, and Mohsen Pourahmadi of the Department of Statistics and Professor Sue Geller of the Department of Mathematics.

The prostate data analyzed in Chapter 2 and Chapter 3 were obtained from Professor Brad Efron's website and the baseball data in Chapter 2 were prepared by Brown [2008]. All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by fellowship from Texas A&M University and NSF grant DMS-0604801.

NOMENCLATURE

μ_i	Unobserved variable of interest, the mean of the i^{th} data, $i = \ldots, q$.
σ_i^2	Unobserved variable of interest, the variance of the i^{th} data, $i = \ldots, q$.
$\boldsymbol{\mu}$	Vector of true means, $(\mu_1, \ldots, \mu_q)^T$
$\boldsymbol{D}$	Diagonal variance matrix where i^{th} diagonal element is σ_i^2.
ϵ_{ij}	Unobserved scaled error for i^{th} data, j^{th} replicate.
X_{ij}	j^{th} replicate of the unobserved μ_i, can be written as $\mu_i + \sigma_i \epsilon_{ij}$, $j = 1 \ldots, n$, $i = \ldots, q$.
$\boldsymbol{X}_{i\cdot}$	Vector of length n, $(X_{i1}, \ldots, X_{in})^T$.
$\boldsymbol{X}_{\cdot j}$	Vector of length q, $(X_{1j}, \ldots, X_{qj})^T$.
$\mathbf{X}$	All $q \times n$ observations, $\boldsymbol{X}_{1\cdot}, \ldots, \boldsymbol{X}_{q\cdot}$.
$\bar{X}_{i\cdot}$	i^{th} sample mean, $n^{-1} \sum_{j=1}^{n} X_{ij}$.
$S_{i\cdot}^2$	i^{th} sample variance, $(n-1)^{-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{i\cdot})^2$.
$X_{i(1)}$	Sample minimum of i^{th} data, $\boldsymbol{X}_{i\cdot}$.
$X_{i(n)}$	Sample maximum of i^{th} data, $\boldsymbol{X}_{i\cdot}$.
$\overline{\boldsymbol{X}}$	Vector of sample means, $(\bar{X}_{1\cdot}, \ldots, \bar{X}_{q\cdot})^T$.
$\boldsymbol{S}^2$	Vector of sample variances, $(S_{1\cdot}^2, \ldots, S_{q\cdot}^2)^T$.
$\mathrm{IQR}(\boldsymbol{y})$	Interquartile range of vector $\boldsymbol{y} = (y_1, \ldots, y_k)^T$.
$\min(\boldsymbol{y})$	Minimum of vector $\boldsymbol{y} = (y_1, \ldots, y_k)^T$.
$\max(\boldsymbol{y})$	Maximum of vector $\boldsymbol{y} = (y_1, \ldots, y_k)^T$.
$\boldsymbol{\delta}$, $\widehat{\boldsymbol{\mu}}$	Vector estimate of the true mean vector, $\boldsymbol{\mu}$.

$\widehat{\boldsymbol{D}}$	Matrix estimate of the true variance matrix, $\boldsymbol{D}$.
$\|\boldsymbol{y}\|^2$	Squared Euclidean norm of vector $\boldsymbol{y} = (y_1, \ldots, y_k)^T$, $\sum_{l=1}^{k} y_l^2$.
$L(\boldsymbol{\delta}, \boldsymbol{\mu})$	Average loss of estimating $\boldsymbol{\mu}$ with $\boldsymbol{\delta}$, $q^{-1}\|\boldsymbol{\delta} - \boldsymbol{\mu}\|^2$, This is function of $\boldsymbol{\mu}$, $\overline{\boldsymbol{X}}$, and $\boldsymbol{D}$.
$R(\boldsymbol{\delta}, \boldsymbol{\mu})$	Average risk of estimating $\boldsymbol{\mu}$ with $\boldsymbol{\delta}$, $E_{\boldsymbol{\mu}} L(\boldsymbol{\delta}, \boldsymbol{\mu})$, This is function of $\boldsymbol{\mu}$ and $\boldsymbol{D}$.
$\boldsymbol{I}$	Identity matrix of appropriate order.
f_ϵ	Probability density function for random variable ϵ.
f_μ	Probability density function for random variable μ.
f_{σ^2}	Probability density function for random variable σ^2.
f_{μ,σ^2}	Joint probability density function for random variable (μ, σ^2).
$U(a, b)$	Uniform density between a and b.
$N(a, b)$	Normal density with mean a and variance b.
$N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	q-dimensional Normal density with mean vector $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.
$G(a, b)$	Gamma density with shape parameter a and rate parameter b.
$IG(a, b)$	Inverse-gamma density with shape parameter a and rate parameter b.
$D(\boldsymbol{\alpha})$	Dirichlet distribution with concentration parameter $\boldsymbol{\alpha}$.
$N\Gamma^{-1}(m, \lambda, a, b)$	Normal-inverse gamma density function with parameters (m, λ, a, b).
χ_k^2	Chi-square density with degree of freedom k.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Estimating the Mean and Variance of a Multivariate Normal Distribution

An old and simple problem in statistics involves estimating the mean of a normal distribution. A somewhat newer and more complex problem is that of estimating the means of many normal distributions simultaneously when we observe independent samples from these distributions. We consider a version of the latter problem in which $X_{i1}, \ldots, X_{in}$ are observations from a normal distribution with mean μ_i and σ_i^2, for $i = 1, \ldots, q$. This can be written as

$$X_{ij} \sim N(\mu_i, \sigma_i^2), \quad j = 1, \ldots, n, \quad i = 1, \ldots, q, \tag{1.1}$$

where $N(a, b)$ denotes a normal density with mean a and variance b. The main goal is to estimate (μ_i, σ_i^2), $i = 1, \ldots, q$, from X_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, q$.

If $\sigma_1^2, \ldots, \sigma_q^2$ are known, to estimate μ_i, n can be as small as 1. We may assume $n = 1$, in which case model (1.1) reduces to

$$X_{i1} \sim N(\mu_i, \sigma_i^2), \quad i = 1, \ldots, q. \tag{1.2}$$

In this case, we observe the pairs (X_{i1}, σ_i^2), $i = 1, \ldots, q$, and the main goal is to estimate the unknown parameters μ_i, $i = 1, \ldots, q$.

In multivariate notation, $\boldsymbol{X}_{\cdot j} = (X_{1j}, \ldots, X_{qj})^T$, $j = 1, \ldots, n$, are n observations from a q-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^T$ and variance $\boldsymbol{D} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_q^2)$, denoted by $N_q(\boldsymbol{\mu}, \boldsymbol{D})$.

In one-dimensional framework, i.e. $q = 1$, the sample mean, $\bar{X}_{1\cdot} = n^{-1} \sum_{j=1}^{n} X_{1j}$, and $(n+1)^{-1} \sum_{i=1}^{n} (X_{1j} - \bar{X}_{1\cdot})^2$ are optimal mean squared error estimators of the population mean and variance, respectively. However, this result does not extend to high-dimensions,

as Stein [1956] showed that the sample means are inadmissible when $q \geq 3$. The seminal work of James and Stein [1961] showed that shrinkage estimators of the means perform better than sample means in terms of mean squared error when $q \geq 3$ and $\sigma_1^2, \ldots, \sigma_q^2$ are all the same (the homoscedastic case) and known. A nice introduction of this class of estimators can be found in the book of Efron [2012]. Efron and Morris [1973] gave an empirical Bayes interpretation of this shrinkage estimator and developed several competing estimators. They noted that even when all σ_i^2 are known, the James-Stein estimator cannot be extended under heteroscedasticity by simply using the transformation $\sigma_i^{-1} X_{ij}$. This is because the shrinkage factor remains constant under the transformation, as opposed to what intuition entails, namely that more shrinkage should be applied to the components with larger σ_i^2. They assumed a hierarchical normal model in which $\mu_i \overset{i.i.d}{\sim} N(0, A)$, and estimated the variance A from the marginal density of X_{ij}. As noted by Efron and Morris [1973], such a hierarchical model is a "Bayesian statement of belief that the μ_i are of comparable magnitude," a belief which is not always realistic.

There is a large literature on estimating the means of a multivariate normal distribution under homoscedasticity, using both frequentist and Bayesian approaches. For example, Baranchik [1970] derived the general form of a minimax estimator for the homoscedastic case. Brown [1971] derived a general condition for Bayes estimators to be admissible in terms of mean squared error. Using these conditions, Berger and Strawderman [1996] showed that some common choices of improper prior on hyperparameters lead to inadmissible estimators, and encouraged the use of a proper prior on hyperparameters. Brown and Greenshtein [2009] proposed a nonparametric empirical Bayes solution for estimating the mean.

In contrast, the literature on the heteroscedastic case is scant. Berger [1976] provided a minimax estimator when the covariance matrix $\boldsymbol{D}$ is known under general quadratic loss. However, this estimator exhibits the counter-intuitive behavior mentioned before. Recently, there have been a few articles addressing this issue. Xie et al. [2012] assumed that $\boldsymbol{D}$ is known and estimated the mean vector $\boldsymbol{\mu}$ using Stein's unbiased risk estimator (SURE).

They showed that the empirical Bayes maximum likelihood estimator (EBMLE) and SURE do not provide the same solution as in the homoscedastic case and proved a few results about the consistency of the SURE. By not limiting the prior on the normal density, they explored a semiparametric option which we will discuss in detail later. Jing et al. [2016] further extended the work of Xie et al. [2012] in the heteroscedastic case when $\boldsymbol{D}$ is unknown by modifying the loss function and assuming a gamma prior on the precision parameter, the inverse of the variance parameter.

Theorem 5.7 of Lehmann and Casella [2006] provided a condition for which the shrinkage estimator becomes a minimax estimator under squared error loss. However, the family of estimators that were considered applies constant shrinkage to all coordinates, as opposed to the common intuition referred to before. Tan et al. [2015] proposed a minimax estimator when the covariance matrix $\boldsymbol{D}$ is known under arbitrary quadratic loss, where the shrinking direction is open to specification and the shrinking factor is determined. This minimax estimator is similar to the estimator arising from the assumption that $\mu_1, \ldots, \mu_q$ are independent with $\mu_i \sim N(0, A_i)$, $i = 1, \ldots, q$. Zhang and Bhattacharya [2017] developed an empirical Bayes method to estimate a sparse normal mean. Weinstein et al. [2018] developed an empirical Bayes estimator assuming that $\sigma_1^2, \ldots, \sigma_q^2$ are part of the random observations. They binned the pairs (X_i, σ_i^2) on the basis of σ_i^2 and applied a spherically symmetric estimator separately in each group. Even though we also assume that (μ_i, σ_i^2) come from a joint distribution, f_{μ,σ^2}, our method is based on modeling the bivariate density of (μ, σ^2) with a flexible mixture of normal-inverse gamma densities and then estimating $\boldsymbol{\mu}$ and $\boldsymbol{D}$.

1.1.1 Motivation for a New Estimator

1.1.1.1 Homoscedastic Case

Consider (1.1), where $\sigma_i^2 = \sigma^2$, for $i = 1, \ldots, q$, and σ^2 is known. We will discuss some existing approaches to estimating $\boldsymbol{\mu}$ in this setting and also how our methodology is related to these approaches. Let $\overline{\boldsymbol{X}}$ be the q-vector whose ith component is the sample mean

$\bar{X}_{i\cdot} = n^{-1}\sum_{j=1}^{n} X_{ij}$, $i = 1, \ldots, q$. Then $\overline{\boldsymbol{X}}$ is distributed as $N_q\left(\boldsymbol{\mu}, n^{-1}\sigma^2\boldsymbol{I}\right)$.

James and Stein [1961] considered a class of estimators indexed by c, which are written as

$$\boldsymbol{\delta}_c^{JS}(\overline{\boldsymbol{X}}) = \left(1 - \frac{\sigma^2}{n}\frac{c}{\left\|\overline{\boldsymbol{X}}\right\|^2}\right)\overline{\boldsymbol{X}},$$

where $\left\|\overline{\boldsymbol{X}}\right\|^2 = \overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}}$ and the i^{th} element of $\boldsymbol{\delta}$, δ_i, is an estimator of μ_i. The average loss, defined by $L(\boldsymbol{\delta}, \boldsymbol{\mu}) = q^{-1}\|\boldsymbol{\delta} - \boldsymbol{\mu}\|^2$, is used to compare different estimates. James and Stein [1961] showed that the constant $c = q-2$ minimizes the risk, $R(\boldsymbol{\delta}, \boldsymbol{\mu}) = E_{\boldsymbol{\mu}}L(\boldsymbol{\delta}, \boldsymbol{\mu})$, for every $\boldsymbol{\mu}$ if $q \geq 3$. We shall call the estimator $\boldsymbol{\delta}_{q-2}^{JS}$ simply $\boldsymbol{\delta}^{JS}$. James and Stein [1961] showed that if $q \geq 3$, $\boldsymbol{\delta}^{JS}$ dominates the MLE, $\overline{\boldsymbol{X}}$, in terms of $R(\boldsymbol{\delta}, \boldsymbol{\mu})$ for every choice of $\boldsymbol{\mu}$.

Baranchik [1970] considered the following more general family of estimators:

$$\boldsymbol{\delta}_r^{JS}(\overline{\boldsymbol{X}}) = \left(1 - \frac{\sigma^2}{n}\frac{r\left(\left\|\overline{\boldsymbol{X}}\right\|^2\right)}{\left\|\overline{\boldsymbol{X}}\right\|^2}\right)\overline{\boldsymbol{X}},$$

and showed that the estimator is minimax if $r(\cdot)$ is monotone, non-decreasing, and such that $0 \leq r(\cdot) \leq 2(q-2)$. Chapter 5 of Lehmann and Casella [2006] discusses risk properties of these estimators in detail. Another minimax estimator is a version of the James-Stein estimator with non-negative multiplier:

$$\boldsymbol{\delta}^{JS+}(\overline{\boldsymbol{X}}) = \max\left(0, 1 - \frac{\sigma^2}{n}\frac{q-2}{\left\|\overline{\boldsymbol{X}}\right\|^2}\right)\overline{\boldsymbol{X}}.$$

This estimator dominates the usual James-Stein estimator in terms of $R(\boldsymbol{\delta}, \boldsymbol{\mu})$. All of these shrinkage estimators shrink each coordinate towards 0.

Efron and Morris [1973] showed an empirical Bayes connection with the James-Stein estimator by assuming a prior of the form $\mu_i \overset{i.i.d}{\sim} N(m, \lambda)$, $i = 1, \ldots, q$, where m and

λ are unknown hyperparameters. From Bayes rule, we have $\mu_1, \ldots, \mu_q$ are independent conditioning on $\bar{X}_{1\cdot}, \ldots, \bar{X}_{q\cdot}, m, \lambda$ with

$$\mu_i \sim N\left(\frac{\lambda}{\lambda + \frac{\sigma^2}{n}}\bar{X}_{i\cdot} + \frac{\frac{\sigma^2}{n}}{\lambda + \frac{\sigma^2}{n}}m, \frac{1}{\lambda^{-1} + \frac{n}{\sigma^2}}\right), \quad i = 1, \ldots, q, \tag{1.3}$$

which leads to the shrinkage estimator

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} - \frac{\frac{\sigma^2}{n}}{\lambda + \frac{\sigma^2}{n}}\left(\overline{\boldsymbol{X}} - m\right).$$

This estimator is a function of the unknowns m and λ. These parameters may be estimated using the marginal density

$$\bar{X}_{i\cdot}|m, \lambda \overset{i.i.d}{\sim} N\left(m, \lambda + \frac{\sigma^2}{n}\right), \quad i = 1, \ldots, q,$$

from which one may obtain the maximum likelihood estimator (MLE) or method of moments estimator (MOM) of (m, λ).

1.1.1.2 Heteroscedastic Case

When $\sigma_1^2, \ldots, \sigma_q^2$ are not all the same but known, we can modify the James-Stein estimator by using the transformation $\sigma_i^{-1}X_{ij}$, which produces homoscedastic data. Then the James-Stein estimate of $\boldsymbol{\mu}$ is

$$\boldsymbol{\delta}^{JS}(\overline{\boldsymbol{X}}) = \left(1 - \frac{q-2}{n\sum_{i=1}^{q}\left(\frac{\bar{X}_{i\cdot}}{\sigma_i}\right)^2}\right)\overline{\boldsymbol{X}}.$$

As discussed in Efron and Morris [1973], this estimate is not intuitive as we should shrink more those coordinates with larger σ_i^2.

When $\sigma_1^2, \ldots, \sigma_q^2$ are not all the same but known, then by assuming the same normal

prior that leads to (1.3) we obtain

$$\mu_i|\bar{X}_{i\cdot}, m, \lambda \sim N\left(\frac{\lambda}{\lambda+\frac{\sigma_i^2}{n}}\bar{X}_{i\cdot}+\frac{\frac{\sigma_i^2}{n}}{\lambda+\frac{\sigma_i^2}{n}}m, \frac{1}{\lambda^{-1}+\frac{n}{\sigma_i^2}}\right), \quad i=1,\ldots,q, \tag{1.4}$$

which leads to the shrinkage estimator $\overline{\boldsymbol{X}} - \boldsymbol{S}\left(\overline{\boldsymbol{X}} - m\right)$, where $\boldsymbol{S}$ is a diagonal matrix with i^{th} diagonal element equal to $\frac{\sigma_i^2}{n}\left(\lambda+\frac{\sigma_i^2}{n}\right)^{-1}$. To estimate the unknown hyperparameters m and λ, we may use the marginal density,

$$X_{ij}|m,\lambda \sim N\left(m, \lambda+\sigma_i^2\right), \quad \text{independently}, \quad \text{for} \quad j=1,\ldots,n, \quad i=1\ldots,q.$$

However, unlike the homoscedastic case, we cannot estimate λ consistently from this marginal density (with n fixed), which impairs the traditional empirical Bayes approach.

Xie et al. [2012] addressed this issue and used the SURE, which finds a solution of m and λ by minimizing an unbiased estimator of the risk $R(\boldsymbol{\delta}, \boldsymbol{\mu})$. They showed that the SURE are optimal in an asymptotic sense compared to EBMLE or EBMOM. To generalize the estimate, they developed a novel semiparametric approach by not assuming a normal-normal hierarchical model. The semiparametric SURE shrinkage estimation which was discussed in Xie et al. [2012] assumed that

$$\hat{\mu}_i^{SM} = (1-b_i)\bar{X}_{i\cdot} + b_i m, \quad i=1,\ldots,q. \tag{1.5}$$

The unbiased estimator of the risk for this estimator is

$$SURE^{SM}(\boldsymbol{b}, m) = q^{-1}\sum_{i=1}^{q}\left(b_i^2\left(\bar{X}_{i\cdot}-m\right)^2 + (1-2b_i)\frac{\sigma_i^2}{n}\right),$$

where $\boldsymbol{b} = (b_1, \cdots, b_q)$. The estimator of $\boldsymbol{b}$ and m is

$$(\hat{\boldsymbol{b}}, \hat{m}) = \arg\min_{\boldsymbol{b},m} SURE^{SM}(\boldsymbol{b}, m),$$

subject to

$$0 \leq b_i \leq 1,\ i = 1, \ldots, q, \quad \text{and} \quad b_i \leq b_j \text{ for any } i \text{ and } j \text{ such that } \frac{\sigma_i^2}{n} \leq \frac{\sigma_j^2}{n}.$$

In principle, all of $b_1, \ldots, b_q$ can be distinct if $\sigma_i^2/n \leq (\bar{X}_{i\cdot} - m)^2$, $i = 1, \ldots, q$, and $\frac{\sigma_i^2}{n\left(\bar{X}_{i\cdot} - m\right)^2} \leq \frac{\sigma_j^2}{n\left(\bar{X}_{j\cdot} - m\right)^2}$ in all cases where $\frac{\sigma_i^2}{n} \leq \frac{\sigma_j^2}{n}$. If these conditions do not hold, the number of distinct b_i reduces. In practice, the number of distinct b_i is very low compared to q since

$$\text{Prob}\left(\sigma_i^2(\bar{X}_{j\cdot} - m)^2 > \sigma_j^2(\bar{X}_{i\cdot} - m)^2\right)$$

is often relatively large even if $\sigma_i^2 < \sigma_j^2$. A natural extension of SURE minimization where all of $m_1, \ldots, m_q$ are distinct is not possible because the solution will be $m_i = \bar{X}_{i\cdot}$, i.e. $b_i = 0$, leading to a non-shrinkage estimator.

The approach of Xie et al. [2012] is tantamount to assuming that $\mu_1, \ldots, \mu_q$ are drawn from a mixture of normals that are all centered at m but have different variances. This is less general than the approach considered in the current paper where we consider a mixture distribution whose components can have different means *and* variances.

Weinstein et al. [2018] proposed a group-linear empirical Bayes method, which treats known variances as part of the random observations and applies a spherically symmetric estimator to each group separately. This shrinks $\bar{X}_{1\cdot}, \ldots, \bar{X}_{q\cdot}$ in different directions, but their clustering mechanism only depends on σ_i^2. This is unrealistic as the shrinkage directions should depend on the modes of the distributions of the unobserved μ_i, and the shrinkage factors should depend on the known σ_i^2. If μ_i is a smooth function of σ_i^2, group linear algorithms perform well as the clustering by similar $\log(\sigma_i^2)$ means unobserved values of μ_i in the same cluster are also similar. However, if μ_i and σ_i^2 are independent, clustering by group linear algorithms is not effective, resulting in poor estimates compared to SURE methods.

Weinstein et al. [2018] obtained results for the heteroscedastic case where $\sigma_1^2, \ldots, \sigma_q^2$ are

i.i.d. Likewise, our proposed method assumes that $\sigma_1^2, \ldots, \sigma_q^2$ are i.i.d., but it has at least two practical advantages over that of Weinstein et al. [2018]. First of all, we need not assume that $\sigma_1^2, \ldots, \sigma_q^2$ are known, and secondly no binning of $\sigma_1^2, \ldots, \sigma_q^2$ (with the attendant problem of choosing the number of bins) is required. We model the joint density of (μ_i, σ_i^2) by a flexible mixture of normal-inverse gamma distributions. As we will show later, our estimators of μ_i are similar in form to the SURE estimates in (1.5), but, when appropriate, they shrink $\bar{X}_{i\cdot}$ towards the mean of a mixture component rather than towards the overall mean. This has the potential of producing better estimates of $\mu_1, \ldots, \mu_q$ when the distribution of μ_i is different from normal.

Jing et al. [2016] extended the result from Xie et al. [2012] to the case where $\sigma_1^2, \ldots, \sigma_q^2$ are unknown. They used a different risk function, $q^{-1} \sum_{i=1}^{q} E_{\boldsymbol{\mu},\boldsymbol{D}} \left((\hat{\mu}_i - \mu_i)^2 + n^{-2} (\hat{\sigma}_i^2 - \sigma_i^2)^2 \right)$, and then minimized unbiased estimators of it by shrinking sample mean and sample variance, $\bar{X}_{i\cdot}$ and $S_{i\cdot}^2$ respectively, towards appropriate direction, where $S_{i\cdot}^2 = (n-1)^{-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{i\cdot})^2$. However, they used constant shrinkage factors for estimating each of μ_i and σ_i^2. Our method naturally extends to the case where $\sigma_1, \ldots, \sigma_q^2$ are unknown.

In Chapter 2, $\boldsymbol{\mu}$ and $\boldsymbol{D}$ will be estimated using a $N\Gamma^{-1}$mixture, which is a more flexible prior than using a single normal. Each $N\Gamma^{-1}$component has a different mean and we shrink each μ_i in an appropriate direction rather than one general direction, which was a main drawback in all previous works.

1.2 Location-Scale Density Estimation in a Random Effects Model

Another common problem in modern statistics is estimating the density of a random variable that is observed with error. Suppose that one observes a few replicates of the true variable with additive measurement error having an unknown density. In principle, the measurement error densities for different sets of replicates could differ in an arbitrary manner. It is reasonable however to assume that the measurement error densities have some degree of commonality. We assume that the measurement error densities are normal with scales that vary with the values of the true variable. A possible model for such data follows:

We observe X_{ij}, the j^{th} replicate of the unobserved variable value μ_i, where

$$\begin{aligned} X_{ij} &= \mu_i + \sigma_i \epsilon_{ij}, \quad i = 1, \ldots, q, \quad j = 1, \ldots, n, \\ \epsilon_{ij} &\sim f_\epsilon \quad i = 1, \ldots, q, \quad j = 1, \ldots, n. \end{aligned} \tag{1.6}$$

The following assumptions are made:

(i) The unknown pairs (μ_i, σ_i^2), $i = 1, \ldots, q$, are independent and identically distributed and follow an unknown absolutely continuous distribution with density f_{μ,σ^2}.

(ii) The unobserved errors ϵ_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, n$, are independent and identically distributed as f_ϵ, which is a known density with mean 0 and variance 1.

(iii) The parameters (μ_i, σ_i^2), $i = 1, \ldots, q$, are independent of ϵ_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, n$.

One goal of this dissertation is to explore different ways of estimating the unknown density f_{μ,σ^2}.

We refer to model (1.6) as the location-scale random effects (LSRE) model. If $\sigma_i^2 = \sigma^2$, for $i = 1, \ldots, q$, then (1.6) reduces to the location random effects (LRE) model. Such models have been used in microarray analyses where X_{ij} is the expression (or log-expression) level for the i^{th} gene of the j^{th} individual. In the LSRE model, the distributions of the small datasets differ only with respect to location and scale. If n is fixed and $q \to \infty$, a bivariate kernel density estimator using $(\bar{X}_{i\cdot}, S_{i\cdot}^2)$, $i = 1, \ldots, q$, where $\bar{X}_{i\cdot} = n^{-1} \sum_{j=1}^n X_{ij}$ and $S_{i\cdot}^2 = (n-1)^{-1} \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot})^2$, is an inconsistent estimator of f_{μ,σ^2}. This thesis proposes a Bayesian semiparametric approach for robust estimation of the density f_{μ,σ^2}, which to our knowledge has not been previously considered.

Literature on the density deconvolution problem dates back at least to 1950. Reiersøl [1950] and Wolfowitz [1957] together proved that, under some general conditions, in the LRE model f_μ and f_ϵ are both nonparametrically estimable when $n \geq 2$. Most of the early literature on density deconvolution used Fourier transform methods to deal with non-replicated

measurements and assumed that measurement errors are independently and identically distributed with known density and constant variance (LRE model). Not much work had been done on the LRE model when both f_μ and f_ϵ are unknown. Much of this literature is referenced by Carroll and Hall [2004]. Of course, in reality a known f_ϵ and constant variance are both very strong assumptions and violation of these assumptions may create bias in the estimation. More recent literature relaxed assumptions on the error density and considers replicated observations. This literature includes the articles Horowitz and Markatou [1996], Li and Vuong [1998], Carroll and Hall [2004], Lin and Carroll [2006], Delaigle et al. [2008], McIntyre and Stefanski [2011], and Hart and Cañette [2011], all of which assumed that f_ϵ is known and used replicated observations in the LRE model. Many of these articles, including Horowitz and Markatou [1996], Carroll and Hall [2004], Lin and Carroll [2006], and Hall and Ma [2007], assumed that f_ϵ is symmetric. Delaigle and Hall [2016] worked with non-replicated observations in the LRE model and assumed that the shape of f_ϵ is unknown but symmetric and then estimated the densities f_μ and f_ϵ nonparametrically.

All the literature mentioned above focuses mainly on the LRE model. Relatively little work had been done on the heteroscedastic error (LSRE model). Staudenmayer et al. [2008] relaxed the assumption of homoscedasticity and worked with the LSRE model. They used Bayesian methodology and modeled ${\sigma_i}^2$ with a variance function that depends on μ_i using a penalized positive mixture of normalized quadratic B-splines. The scaled measurement errors were assumed to be normally distributed. Hart and Cañette [2011] proposed a minimum distance estimator to obtain nonparametric estimates of the distributions without assuming that f_ϵ is symmetric in the LSRE model. They also formulated a distribution-free rank test of the LRE model against the LSRE model when $n \geq 4$. Sarkar et al. [2014] used Bayesian methods to model f_μ by a location-scale mixture of normals induced by a Dirichlet process. The scaled error distribution f_ϵ is more flexible and in our approach will be modeled by an infinite mixture model induced by a Dirichlet process.

In Chapter 3, f_{μ,σ^2} will be estimated using a $N\Gamma^{-1}$mixture. However, in this chapter

we assume that f_ϵ be a standard normal. In Chapter 4, f_{μ,σ^2} will be estimated using a bivariate histogram when the scaled error distribution, f_ϵ, is any known density with mean 0 and variance 1. We do not assume any functional dependency between μ and σ^2 in both chapters.

2. ESTIMATING THE MEAN AND VARIANCE OF A NORMAL VECTOR

As mentioned in Section 1.1, the main focus of this chapter is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{D}$ in model (1.1) and (1.2). To achieve that first we need to estimate the joint density of (μ, σ^2), f_{μ,σ^2}, using $N\Gamma^{-1}$mixture which we will discuss in Section 2.1 and 2.2.

2.1 Modeling the Joint Distribution of Location-Scale

We define gamma and inverse-gamma densities as

$$G(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} I_{(0,\infty)}(x) \text{and} \quad IG(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x} I_{(0,\infty)}(x), \tag{2.1}$$

respectively, where Γ is the gamma function and I_A is the indicator function defined as

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise.} \end{cases}$$

Though it is more common to use a mixture of normal densities, σ^2 has support only on the positive side of the real line, and hence using a mixture of bivariate normals seems unreasonable. An easy way to get around the problem of positive support is to estimate the density of $\log(\sigma^2)$ using a mixture of normals. However, if we assume f_ϵ is standard normal, then a mixture of bivariate normals for the joint density of $(\mu, \log(\sigma^2))$ is not a conjugate prior. A mixture of normal-inverse-gamma ($N\Gamma^{-1}$) densities leads to a posterior density belonging to a known family of densities. A $N\Gamma^{-1}(m, \lambda, \alpha, \beta)$ density has two components, normal and inverse-gamma, and is defined by

$$g(\mu, \sigma^2|m, \lambda, \alpha, \beta) = N(\mu|m, \sigma^2/\lambda) IG(\sigma^2|\alpha, \beta).$$

The density f_{μ,σ^2} is defined to be a mixture of $N\Gamma^{-1}$ densities induced by a Dirichlet

process with concentration parameter γ. Let $\boldsymbol{\pi}$ denote the vector of random mixture weights. Sethuraman [1994] describes the *stick-breaking* process, a method to construct $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty}$ so that $\sum_{k=1}^{\infty} \pi_k = 1$. For $r = 1, 2, \ldots$, the process is such that

$$\pi_r = s_r \prod_{j=1}^{r-1}(1 - s_j), \quad s_1, s_2, \ldots \overset{i.i.d}{\sim} \text{Beta}(1, \gamma),$$

which we denote $D(\gamma)$. The quantities $\boldsymbol{m}$, $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are the vectors of parameters of the $N\Gamma^{-1}$ densities that make up the mixture. Let $\boldsymbol{\Theta} = [\boldsymbol{m}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ be a matrix of four columns whose r^{th} row, $\boldsymbol{\Theta}_r$, contains parameters for the r^{th} component of the mixture. The Dirichlet process mixture model (DPMM), denoted $DP(\gamma, G_0)$ with concentration parameter γ, base measure G_0 and $N\Gamma^{-1}$ mixture components, is specified as

$$f_{\mu,\sigma^2}(\mu, \sigma^2|\boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{r=1}^{\infty} \pi_r g(\mu, \sigma^2|\boldsymbol{\Theta}_r), \quad \boldsymbol{\Theta}_r \overset{i.i.d}{\sim} G_0(\,\cdot\,|\boldsymbol{\Theta_H}), \quad \boldsymbol{\pi} \sim D(\gamma).$$

The distribution $G_0(\,\cdot\,|\boldsymbol{\Theta_H})$ depending on parameters $\boldsymbol{\Theta_H} = (m_0, \zeta^2, a_\lambda, b_\lambda, a_\alpha, b_\alpha, a_\beta, b_\beta)$ is the prior for the component parameters and is taken to be as follows: m_r, λ_r, α_r, and β_r are independent with

$$m_r \sim N(m_0, \zeta^2), \quad \lambda_r \sim G(a_\lambda, b_\lambda), \quad \alpha_r \sim G(a_\alpha, b_\alpha), \quad \beta_r \sim G(a_\beta, b_\beta).$$

Even though the mixture model theoretically has a countably infinite number of components, given a data set, one can only use a mixture model with a finite number of components. Indeed, in practice, a finite number of components is adequate. Ishwaran and James [2001] constructed a useful class of truncated Dirichlet processes, denoted $DP_k(\gamma, G_0)$, by applying truncation to standard Dirichlet processes, where the number of components is fixed at k. The truncation is applied by assuming $\pi_{k+1} = \pi_{k+2} = \cdots = 0$ and replacing π_k by $1 - \sum_{r=1}^{k-1} \pi_r$. They showed that the expected sum of moments of discarded random weights decreases exponentially fast in k, and thus, for a moderate k, we should be able to achieve

an accurate approximation. We shall use $DP_k(\gamma, G_0)$ in order to model the density f_{μ,σ^2}.

Since we have measurement error, we do not observe the pair (μ_i, σ_i^2) directly. Instead, we observe $\{X_{ij}\}_{j=1}^n$, which will be referred to as $\boldsymbol{X}_{i\cdot}$, a vector of observed replications of the true unobserved variable μ_i. As we already assumed the error density to be standard normal, the joint density of $\boldsymbol{X}_{i\cdot}$ given μ_i and σ_i^2 is

$$f(\boldsymbol{X}_{i\cdot}|\mu_i, \sigma_i^2) = \prod_{j=1}^{n} \frac{1}{\sigma_i} f_\epsilon \left(\frac{X_{ij} - \mu_i}{\sigma_i} \right) = \prod_{j=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2\sigma_i^2}(X_{ij}-\mu_i)^2}.$$

Let Z_i be a latent variable indicating the component of the mixture distribution from which the pair (μ_i, σ_i^2) was drawn. The conditional joint density of (μ_i, σ_i^2) is

$$f(\mu_i, \sigma_i^2|\boldsymbol{\Theta}, Z_i = z_i) = g(\mu_i, \sigma_i^2|\boldsymbol{\Theta}_{\boldsymbol{z_i}}).$$

The prior probability mass function (p.m.f.) of the latent variable Z_i is

$$\text{Prob}(Z_i = z_i|\boldsymbol{\pi}) = \pi_{z_i}.$$

Let $U \perp\!\!\!\perp V|W$ denotes that two random variables U and V are independent conditional on W. Let $\mathbb{U}_r = \{i : Z_i = r\}$ and c_r be the cardinality of $\mathbb{U}_r$. Let $\mathbf{X}$ denotes the all $q \times n$ observations, $\boldsymbol{X}_{1\cdot}, \ldots, \boldsymbol{X}_{q\cdot}$. We make the following assumptions:

(i) $\mathbf{X} \perp\!\!\!\perp Z_1, \ldots, Z_q, \boldsymbol{\Theta}, \boldsymbol{\pi}|\boldsymbol{\mu}, \boldsymbol{D}$,

(ii) The conditional distribution of $\boldsymbol{\mu}, \boldsymbol{D}$ given $\boldsymbol{\Theta}, \boldsymbol{\pi}, Z_1 = z_1, \ldots, Z_q = z_q$ is

$$\prod_{i=1}^{q} g(\mu_i, \sigma_i^2|\boldsymbol{\Theta}_{z_i}),$$

(iii) $Z_1, \ldots, Z_q \perp\!\!\!\perp \boldsymbol{\Theta}|\boldsymbol{\pi}$,

(iv) $\boldsymbol{\Theta} \perp\!\!\!\perp \boldsymbol{\pi}$.

The posterior is proportional to

$$
\begin{aligned}
&f(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{D}, z_1, \ldots, z_q, \boldsymbol{\Theta}, \boldsymbol{\pi}) f(\boldsymbol{\mu}, \boldsymbol{D}, z_1, \ldots, z_q, \boldsymbol{\Theta}, \boldsymbol{\pi}) \\
=&\prod_{i=1}^{q} f(\boldsymbol{X}_{i\cdot}|\mu_i, \sigma_i^2) \prod_{i=1}^{q} g(\mu_i, \sigma_i^2|\boldsymbol{\Theta}_{z_i}) \prod_{r=1}^{k} {\pi_r}^{c_r} \prod_{r=1}^{k} G_0(\boldsymbol{\Theta_r}|\boldsymbol{\Theta_H}) D(\boldsymbol{\pi}|\gamma).
\end{aligned}
$$

We may reparametrize α_r and β_r in terms of location and scale parameters. If δ_r denotes a point between the mean and mode of the $IG(\alpha_r, \beta_r)$, then we can rewrite the rate parameter β_r as $\delta_r \alpha_r$. The quantities δ_r and α_r can be treated as location and scale parameters respectively. Since δ_r is the location parameter of a density with positive support, we can use a gamma prior on δ_r just as we did for β_r with shape parameter a_δ and scale parameter b_δ.

2.2 Algorithm to Estimate Unknown Parameters

We will find estimates of the parameters $(\boldsymbol{\Theta}, \boldsymbol{\pi})$ by using an MCMC algorithm to approximate their posterior density. In the notation that follows, $\theta|\cdot$ stands for the conditional distribution of θ given the data and all unknowns besides θ. The full conditional posterior densities of μ_i and σ_i^2 are normal and inverse-gamma, respectively, with the following parameters:

$$
\begin{aligned}
&\mu_i|\cdot \sim N\left(\frac{n\bar{X}_{i\cdot} + m_{z_i}\lambda_{z_i}}{n + \lambda_{z_i}}, \frac{\sigma_i^2}{n + \lambda_{z_i}}\right), \\
&\sigma_i^2|\cdot \sim IG\left(\frac{n+1}{2} + \alpha_{z_i}, \frac{1}{2}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2 + \frac{\lambda_{z_i}}{2}(\mu_i - m_{z_i})^2 + \beta_{z_i}\right), \quad i = 1, \ldots, q.
\end{aligned} \tag{2.2}
$$

Letting $\text{Prob}(z|\cdot)$ denote the posterior p.m.f. of the latent variable Z_i given the data and all other unknowns, we have

$$
\text{Prob}(Z_i = z_i|\cdot) = \frac{\pi_{z_i} g\left(\mu_i, \sigma_i^2|m_{z_i}, \lambda_{z_i}, \alpha_{z_i}, \beta_{z_i}\right)}{\sum_{r=1}^{k} \pi_r g\left(\mu_i, \sigma_i^2|m_r, \lambda_r, \alpha_r, \beta_r\right)} \quad i = 1, \ldots, q.
$$

From expression (2.2) we can interpret λ_{z_i} as a shrinkage parameter. If λ_{z_i} tends to 0 then the posterior density of μ_i is centered at the sample mean. The quantity λ_{z_i} controls the amount of shrinkage towards the mean of the mixture component.

The full conditional posterior densities of the mixture components, $(\boldsymbol{\Theta}, \boldsymbol{\pi})$, are given by

$$m_r|\cdot \sim N\left(\frac{m_0\zeta^{-2} + \lambda_r \sum_{i\in\mathbb{U}_r} \mu_i\sigma_i^{-2}}{\zeta^{-2} + \lambda_r \sum_{i\in\mathbb{U}_r} \sigma_i^{-2}}, \frac{1}{\zeta^{-2} + \lambda_r \sum_{i\in\mathbb{U}_r} \sigma_i^{-2}}\right),\ \lambda_r|\cdot \sim G\left(\frac{c_r}{2} + a_\lambda, \sum_{i\in\mathbb{U}_r} \frac{(\mu_i - m_r)^2}{2\sigma_i^2} + b_\lambda\right),$$

$$\alpha_r|\cdot \sim \frac{{b_\alpha}^{a_\alpha}}{\Gamma(a_\alpha)}{\alpha_r}^{a_\alpha - 1}e^{-b_\lambda\alpha_r} \prod_{i\in\mathbb{U}_r} \frac{{\beta_r}^{\alpha_r}}{\Gamma(\alpha_r)}(\sigma_i^2)^{-\alpha_r - 1}e^{-\beta_r\sigma_i^{-2}},\ \boldsymbol{\pi}|\cdot \sim D\left(c_1 + \frac{\gamma}{k}, \ldots, c_k + \frac{\gamma}{k}\right),$$

$$\beta_r|\cdot \sim \frac{{b_\beta}^{a_\beta}}{\Gamma(a_\beta)}{\beta_r}^{a_\beta - 1}e^{-b_\beta\beta_r} \prod_{i\in\mathbb{U}_r} \frac{(\beta_r)^{\alpha_r}}{\Gamma(\alpha_r)}(\sigma_i^2)^{-\alpha_r - 1}e^{-\beta_r\sigma_i^{-2}}, \quad r = 1, \ldots, k.$$

The full conditional posterior densities of m_r, λ_r, and $\boldsymbol{\pi}$ follow normal, gamma, and Dirichlet densities, respectively. The parameters, α_r and β_r do not have a standard density. We therefore use a Metropolis-Hastings algorithm to sample from these densities.

Denote our estimate of $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}^{DPMM}$, where $DPMM$ stands for the Dirichlet process mixture model. The i^{th} component of $\hat{\boldsymbol{\mu}}^{DPMM}$, $\hat{\mu}_i^{DPMM}$, approximates $E(\mu_i|\text{data})$. Defining

$$\hat{\mu}(z_i, m_{z_i}, \lambda_{z_i}) = (n\bar{X}_{i\cdot} + m_{z_i}\lambda_{z_i})/(n + \lambda_{z_i}),$$

expression (2.2) and iterated expectation imply that

$$E(\mu_i|\text{data}) = E\left[\hat{\mu}(z_i, m_{z_i}, \lambda_{z_i})|\text{data}\right].$$

Letting $b_i = \lambda_{z_i}/(n + \lambda_{z_i})$, we have

$$\hat{\mu}(z_i, m_{z_i}, \lambda_{z_i}) = (1 - b_i)\bar{X}_{i\cdot} + b_i m_{z_i},$$

and so for each choice of the unknown parameters $(z_i, m_{z_i}, \lambda_{z_i})$, $\hat{\mu}(z_i, m_{z_i}, \lambda_{z_i})$ is a shrinkage estimate having the same form as the SURE in (1.5). The actual estimate of μ_i, $E(\mu_i|\text{data})$, is simply the posterior mean of all these shrinkage estimates. In the event that μ_i comes

from, say, component 1 with high probability

$$\hat{\mu}_i^{DPMM} \approx nE((n+\lambda_1)^{-1}|\text{data}, Z_i = 1)\bar{X}_{i\cdot} + E(m_1\lambda_1(n+\lambda_1)^{-1}|\text{data}, Z_i = 1),$$

and hence $\bar{X}_{i\cdot}$ shrinks towards the posterior mean of m_1 rather than the overall mean. Certainly in cases where the distribution of μ_i is multimodal with widely separated modes this scheme should produce much better estimates of $\boldsymbol{\mu}$ than does equation (1.5), a claim confirmed by simulations in Sections 2.4.1-2.4.2.

From equation (2.2) the posterior mean of σ_i^2 is

$$E(\sigma_i^2|\text{data}) = E\left\{ \frac{(n-1)\tilde{\sigma}_i^2 + 2\alpha_{z_i}(\beta_{z_i}/\alpha_{z_i})}{n-1+2\alpha_{z_i}} \middle| \text{data} \right\}, \tag{2.3}$$

where

$$\tilde{\sigma}_i^2 = (n-1)^{-1}\left[(n-1)S_{i\cdot}^2 + n(\bar{X}_{i\cdot} - \mu_i)^2 + \lambda_{z_i}(\mu_i - m_{z_i})^2\right].$$

So, $E(\sigma_i^2|\text{data})$ has an interpretation analogous to that of $E(\mu_i|\text{data})$. The quantity β_{z_i}/α_{z_i} may be regarded as a location parameter of the inverse-gamma component as it lies between the mode and the mean, and therefore $E(\sigma_i^2|\text{data})$ is the posterior mean of shrinkage estimates each of which shrinks the variance estimate $\tilde{\sigma}_i^2$ towards β_{z_i}/α_{z_i}.

2.3 Choice of Prior Parameters

We can run a fully Bayes approach using a prespecified value of $\boldsymbol{\Theta_H}$ and a non-informative prior on $\boldsymbol{\Theta}$, or take an empirical Bayes approach to estimate $\boldsymbol{\Theta_H}$ from the data. Even though we do not observe (μ_i, σ_i^2) directly, we can perceive the problem as one of clustering the (μ_i, σ_i^2) pairs, where each cluster has a different $N\Gamma^{-1}$ density. The parameter m_r denotes the mean of all μ_i that belong to the r^{th} cluster. The parameters m_0 and ζ^2 are the mean and variance of each m_r. Let $\bar{\bar{X}} = (nq)^{-1}\sum_{i=1}^{q}\sum_{j=1}^{n} X_{ij}$ denote the grand mean and $S^2 = (nq-1)^{-1}\sum_{i=1}^{q}\sum_{j=1}^{n}(X_{ij} - \bar{\bar{X}})^2$ the grand variance. It is reasonable to estimate m_0

with its unbiased estimator, the grand mean $\bar{\bar{X}}$. Note that

$$E(X_{ij}|\boldsymbol{\Theta},\boldsymbol{\pi}) = E(E(X_{ij}|\mu_i,\sigma_i^2)|\boldsymbol{\Theta},\boldsymbol{\pi}) = E(\mu_i|\boldsymbol{\Theta},\boldsymbol{\pi}) = \sum_{r=1}^{k}\pi_r m_r,$$
$$E(X_{ij}|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma) = E(E(X_{ij}|\boldsymbol{\Theta},\boldsymbol{\pi})|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma) = E\left(\sum_{r=1}^{k}\pi_r m_r|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma\right) = m_0.$$

On the other hand, estimating ζ^2 is more difficult as the conditional variance of the sample means depends on ζ^2 and many other parameters. Note that

$$\begin{aligned}
\mathrm{var}(\bar{X}_{i\cdot}|Z_i=r,\boldsymbol{\Theta}) =& \mathrm{var}(E(\bar{X}_{i\cdot}|\mu_i,\sigma_i^2)|Z_i=r,\boldsymbol{\Theta}) + E(\mathrm{var}(\bar{X}_{i\cdot}|\mu_i,\sigma_i^2)|Z_i=r,\boldsymbol{\Theta})\\
=& \mathrm{var}(\mu_i|Z_i=r,\boldsymbol{\Theta}) + n^{-1}E(\sigma_i^2|Z_i=r,\boldsymbol{\Theta})\\
=& \frac{\beta_r}{\lambda_r(\alpha_r-1)} + \frac{\beta_r}{n(\alpha_r-1)} = \frac{\beta_r}{(\alpha_r-1)}\left(\frac{1}{n}+\frac{1}{\lambda_r}\right),\\
\mathrm{var}(\bar{X}_{i\cdot}|\boldsymbol{\Theta},\boldsymbol{\pi}) =& \mathrm{var}(E(\bar{X}_{i\cdot}|Z_i=r,\boldsymbol{\Theta}_r)|\boldsymbol{\Theta},\boldsymbol{\pi}) + E(\mathrm{var}(\bar{X}_{i\cdot}|Z_i=r,\boldsymbol{\Theta}_r)|\boldsymbol{\Theta},\boldsymbol{\pi})\\
=& \mathrm{var}(m_r|\boldsymbol{\Theta},\boldsymbol{\pi}) + E\left(\frac{\beta_r}{(\alpha_r-1)}\left(\frac{1}{n}+\frac{1}{\lambda_r}\right)|\boldsymbol{\Theta},\boldsymbol{\pi}\right)\\
=& \sum_{r=1}^{k}\pi_r m_r^2 - \left(\sum_{r=1}^{k}\pi_r m_r\right)^2 + \sum_{r=1}^{k}\frac{\pi_r\beta_r}{(\alpha_r-1)}\left(\frac{1}{n}+\frac{1}{\lambda_r}\right),\\
\mathrm{var}(\bar{X}_{i\cdot}|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma) =& \mathrm{var}(E(\bar{X}_{i\cdot}|\boldsymbol{\Theta},\boldsymbol{\pi})|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma) + E(\mathrm{var}(\bar{X}_{i\cdot}|\boldsymbol{\Theta},\boldsymbol{\pi})|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma)\\
>& \mathrm{var}\left(\sum_{r=1}^{k}m_r\pi_r|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma\right) = \frac{2m_0^2\gamma(k-1)}{\gamma+1} + \frac{\zeta^2(k\gamma+1)}{\gamma+1} \geq \zeta^2.
\end{aligned} \tag{2.4}$$

The inequality in the last line of (2.4) is intuitively clear as ζ^2 can be seen as the between group variance of μ_i, which must be less than the total variance of μ_i. We will use $S^2_{\bar{X}} = q^{-1}\sum_{i=1}^{q}\left(\bar{X}_{i\cdot}-\bar{\bar{X}}\right)^2$ as our choice of ζ^2 in the prior for m_r. Doing so is somewhat informative, but not too informative since $S^2_{\bar{X}}$ estimates $\mathrm{var}(\bar{X}_{i\cdot}|\boldsymbol{\Theta}_{\boldsymbol{H}},\gamma)$, which is larger than ζ^2.

An important parameter of the $N\Gamma^{-1}$ mixtures is λ_r, whose prior has two hyperparameters, a_λ and b_λ. We have

$$\frac{E(\sigma_i^2|Z_i=r,\boldsymbol{\Theta}_r)}{\mathrm{var}(\mu_i|Z_i=r,\boldsymbol{\Theta}_r)} = \lambda_r,$$

which means that λ_r may be regarded as a noise to signal ratio. In many, if not most, cases one anticipates that noise to signal ratios will be smaller than 1, which motivates choosing a_λ and b_λ to produce values of λ_r that are smaller than 1 with fairly high probability.

Similarly, a_α, b_α, a_β, and b_β are the hyperparameters of α_r and β_r, the scale and rate parameters of the inverse-gamma distributions comprising the mixture. We may choose the hyperparameters in such a way that the prior for α_r and β_r has low information.

The prior on mixing probabilities $\boldsymbol{\pi}$ is a Dirichlet density with parameter γ. Ferguson [1983] discussed in detail two independent interpretations of the Dirichlet process parameter γ. The first one concerns the relative size of π_r and the second one concerns prior information. A smaller value of γ means there are big differences in π_r values and also that we mistrust our prior. So, posterior estimates will be strongly influenced by the data.

2.4 Simulation Study

In this section, we conduct a number of simulations to compare different methods of estimating $\boldsymbol{\mu}$ and $\boldsymbol{D}$. We simulated data from either (1.1) or (1.2) using a number of different choices for f_{μ,σ^2}. To evaluate an estimator $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, we approximate the following version of mean squared error:

$$MSE(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = E\left[\frac{1}{q}\sum_{i=1}^{q}(\hat{\mu}_i - \mu_i)^2\right],$$

where the expectation is taken with respect to the joint distribution of $\boldsymbol{X}_{1\cdot}, \ldots, \boldsymbol{X}_{q\cdot}$ given $\boldsymbol{\Theta}, \boldsymbol{\pi}$. In using this risk function we are taking into account randomness due to (μ_i, σ_i^2), $i = 1, \ldots, q$. In our simulation study, each new data set is obtained by generating new values $(\mu_i, \sigma_i^2, \boldsymbol{\epsilon}_{i\cdot})$, $i = 1, \ldots, q$, where $\boldsymbol{\epsilon}_{i\cdot} = (\epsilon_{i1}, \ldots, \epsilon_{in})$. The risk $MSE(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ is then approximated by $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$, the average of $\sum_{i=1}^{q}(\hat{\mu}_i - \mu_i)^2/q$ over all data sets. Similarly, we define $MSE(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ and $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ when we are estimating $\boldsymbol{D}$.

2.4.1 Comparing Different Estimators when Variances are Known

In this section, data are generated from model (1.2) and it is assumed that $\sigma_1^2, \ldots, \sigma_q^2$ are known. Table 2.1 compares $MSE(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ for the methods discussed in Xie et al. [2012] and Weinstein et al. [2018] with our method, denoted $N\Gamma^{-1}$. The estimators of Xie et al. [2012] defined by their expressions (7.1), (7.2), (7.3), (4.2), (5.1), (6.3), and (6.2) will be called EBMLE.XKB, EBMOM.XKB, JS.XKB, SURE.G.XKB, SURE.M.XKB, SURE.SG.XKB, and SURE.SM.XKB, respectively. Weinstein et al. [2018] developed group-linear and dynamic group-linear algorithms, which are referred to here as GL.WMBZ and DGL.WMBZ, respectively. We also consider Oracle.XKB, which, although not an estimate as described in section 7 of Xie et al. [2012], provides a sensible lower bound on a risk estimator with given parametric form. Our estimator does not belong to this class of estimators because the sample means are not shrunk towards a single value, as discussed in Section 2.2.

Examples 1-6 of this section were taken from Xie et al. [2012] and also used by Weinstein et al. [2018]. We simulated data from model (1.2) for different choices of f_{μ,σ^2}. The experiment was repeated 1000 times for each of $q = 20, 60, 100, \ldots, 500$. The resulting values of $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ are shown in Table (2.1) for all q and each of the methods mentioned above.

Example 1. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,1)$ and $\sigma^2 \sim U(0.1,1)$, where $U(a,b)$ denotes the uniform distribution on the interval (a,b). Here and in Examples 2-5, 7, and 8 we take $\epsilon \sim N(0,1)$. Figure 2.1 shows that SURE.M.XKB performs better than SURE.SG.XKB, GL.WMBZ and $N\Gamma^{-1}$ (the only methods plotted) since the generated data conform with the parametric form with equation (1.4) upon which SURE.M.XKB is based. Likewise equation SURE.G.XKB, EBMLE.XKB, and EBMOM.XKB assume that $\boldsymbol{\mu}$ has the parametric form of (1.4), and hence these methods outperform the other methods. Our results (some of which are not given in figure (2.1) or table (2.1)) show that, except for JS.XKB and DGL.WMBZ, all estimated risks converge to the oracle risk. JS.XKB, which applies constant shrinkage for every coordinate results in an inefficient estimator. Interestingly, even though the distribution of σ^2 is uniform, the case where group linear algorithms*

should perform well because of their use of binning, the $N\Gamma^{-1}$ method outperforms the group linear algorithms for small q.

Example 2. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim U(0,1)$ and $\sigma^2 \sim U(0.1,1)$. This example is quite similar to example 1, and shows that the parametric form (1.4) is not necessarily important as long as μ and σ^2 are independent. The estimated risks of EBMLE.XKB, EBMOM.XKB and SURE.M.XKB all converge to the risk of Oracle.XKB. Figure 2.1 shows that SURE.M.XKB and SURE.SG.XKB perform better than the other two methods. The fact that the normal-inverse gamma mixture allows for a dependency between μ and σ^2 may explain why $N\Gamma^{-1}$ does not perform as well as the SURE-based methods. However, $N\Gamma^{-1}$ performs better than GL.WMBZ.*

Example 3. *Here the joint distribution of μ and σ^2 is singular, with $\sigma^2 \sim U(0.1,1)$ and $\mu = \sigma^2$. Rather than being independent, as in examples 1 and 2, μ and σ^2 are highly dependent in this case. Even though the SURE.M.XKB and SURE.SG.XKB risks converge to the Oracle.XKB risk, the Oracle.XKB risk is actually larger than that of GL.WBMZ and $N\Gamma^{-1}$. When μ and σ^2 are dependent, SURE-based methods tend to perform poorly compared to group linear algorithms and $N\Gamma^{-1}$. GL.WBMZ is based on clustering $log(\sigma^2)$, and if μ is a function of σ^2 then group linear algorithms will usually cluster the μ_is correctly, regardless of the distribution of μ. So in this example, group linear methods outperform all the other methods.*

Example 4. *Again the joint distribution of μ and σ^2 is singular with $\mu = \sigma^2$, but now $\frac{1}{\sigma^2} \sim \chi^2_{10}$. The risks of SURE.M.XKB and SURE.SG.XKB converge to that of Oracle.XKB as q increases. The $N\Gamma^{-1}$ method performs better than GL.WMBZ for lower values of q, but as q increases performance of both of these algorithms improves and approaches that of Oracle.XKB.*

Example 5. *In this example, the distribution of σ^2 is discrete and such that σ^2 is either 0.1 or 0.5, each with probability $1/2$, while $\mu|(\sigma^2 = 0.1) \sim N(2, 0.1)$ and $\mu|(\sigma^2 = 0.5) \sim N(0, 0.5)$.*

Obviously, μ and σ^2 are not independent in this case, and there are two distinct groups of data. Both GL.WBMZ and $N\Gamma^{-1}$ effectively treat the two groups separately, whereas SURE.M.XKB and SURE.SG.XKB shrink all means in the same direction, as does Oracle.XKB. For each q, GL.WMBZ and $N\Gamma^{-1}$ greatly outperform the SURE-based methods.

Example 6. *Here the setting is the same as in example 3 except that $\epsilon \sim U(-\sqrt{3}, \sqrt{3})$. As in example 5, for any q, GL.WMBZ and $N\Gamma^{-1}$ outperform the SURE-based methods and GL.WMBZ performs better than $N\Gamma^{-1}$ since μ is a function of σ^2.*

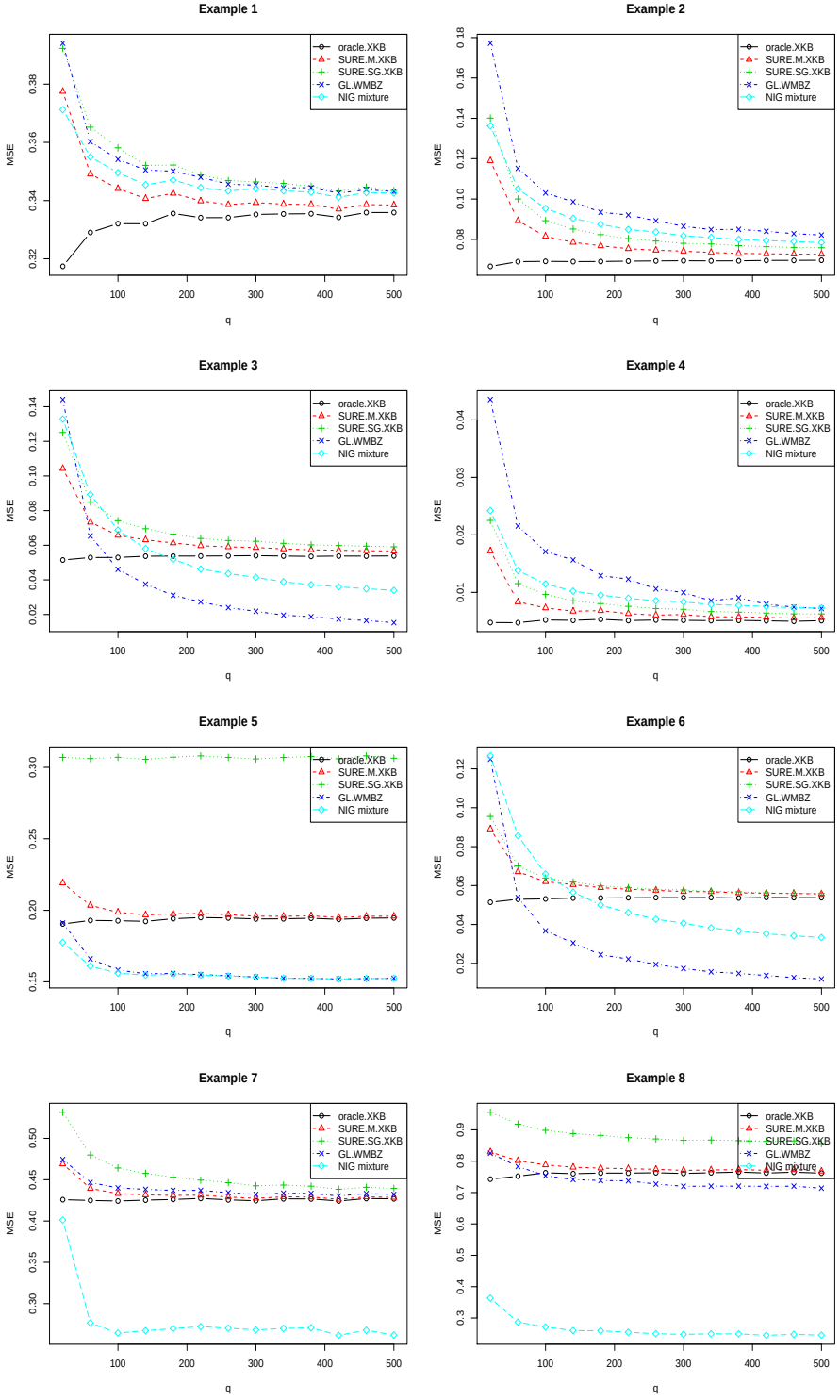
Example 7. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\sigma^2 \sim U(0.1, 1)$ and $\mu \sim 0.5N(0, 0.1) + 0.5N(3, 0.1)$. Here the distribution of μ is bimodal. This is a case where algorithms based on clustering σ^2 fail, and $N\Gamma^{-1}$ does very well. SURE-based methods shrink all X_i in the same direction, towards 1.5, whereas $N\Gamma^{-1}$ shrinks X_i towards either 0 or 3 after identifying the cluster to which μ_i is likely to belong. Group linear methods end up having the same defect in this case as the SURE-based methods. Since clustering is based on $log(\sigma_i^2)$ and μ_i is independent of σ_i^2, each group linear cluster will contain roughly equal numbers of μ_is from the two components. It follows that the group linear algorithms will also shrink X_i towards 1.5.*

Example 8. *The distribution of (μ, σ^2) is such that $(\mu, \sigma^2) \sim N\Gamma^{-1}(2, 2, 5, 2)$ with probability 0.6 and $N\Gamma^{-1}(10, 4, 3, 3)$ with probability 0.4. In this example, the underlying distribution of (μ, σ^2) is a mixture of normal-inverse gammas, and so, as expected, $N\Gamma^{-1}$ method outperforms all the others. As the marginal distribution of μ is bimodal, SURE-based and group linear methods do not perform well for the same reason as in Example 7.*

2.4.2 Comparing Different Estimators when Variances are Unknown

Tables 2.2 and 2.3 compare the different methods discussed in Xie et al. [2012], Weinstein et al. [2018] and Jing et al. [2016]. The method referred to as SURE.M.Double can be found in (11)-(12) of Jing et al. [2016]. Although Jing et al. [2016] discussed a few different double shrinkage algorithms, we have found the performance of those algorithms to be very similar

Figure 2.1: $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ vs. dimension q of normal vector for Examples 1-8 of Section 2.4.1. The dimension sizes are $q = 20, 60, \ldots, 500$ and results are based on 1000 replications at each q.



to each other, and therefore report results only for the algorithm in expression (16) of Jing et al. [2016], which we refer to as SURE.M.Double. As Xie et al. [2012] and Weinstein et al. [2018] assumed that $\sigma_1^2, \ldots, \sigma_q^2$ were known, we do as they suggested and replace σ_i^2 by $S_{i\cdot}^2$ when implementing their algorithms.

We simulated data from model (1.1) for different choices of f_{μ,σ^2}. In all the examples of this section $\epsilon \sim N(0,1)$. For each (μ_i, σ_i^2) pair there are $n = 4$ replications. We only observe X_{ij}, for $i = 1, \ldots, q$, $j = 1, \ldots, n$, and not $\sigma_1^2, \ldots, \sigma_q^2$. We repeat the experiment 1000 times for each q, and $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ and $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ were determined. Tables 2.2 and 2.3 provide estimated risks averaged over all q, and Figure 2.2 shows how our method compares with the two SURE methods discussed in Xie et al. [2012] and with the group linear algorithms discussed in Weinstein et al. [2018]. Figure 2.3 shows how our method of estimating $\boldsymbol{D}$ compares with the SURE.M.Double discussed in Jing et al. [2016].

Example 9. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,3)$ and $\sigma^2 \sim IG(5,2)$. Figure 2.2 shows that our method outperforms the other three when estimating μ_i. Table 2.2 shows that $N\Gamma^{-1}$ performs similarly to the double shrinkage algorithms discussed in Jing et al. [2016]. As the latter algorithms and $N\Gamma^{-1}$ are based on the normal-inverse gamma distribution, and the (μ_i, σ_i^2) distribution, in this case, is normal-inverse gamma, it is not surprising that these methods outperform the others here. Table 2.3 and Figure 2.3 show that the SURE.M.Double method slightly outperforms $N\Gamma^{-1}$ in estimating $\boldsymbol{D}$.*

Example 10. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,3)$ and $\sigma^2 \sim G(9,3)$. This case is similar to Example 7, and likewise the results are similar.*

Example 11. *Here $(\mu, \sigma^2) \sim 0.95N\Gamma^{-1}(2,2,5,2) + 0.05N\Gamma^{-1}(10,4,3,3)$, the same mixture distribution considered in Example 8. This is a case where μ and σ^2 are dependent and their distribution is bimodal. Our algorithm outperforms all other methods in terms of both $\boldsymbol{\mu}$ and $\boldsymbol{D}$ estimation, as seen in Figures 2.2-2.3 and Tables 2.2-2.3.*

Example 12. *In this case μ and σ^2 are independent with $\mu \sim 0.5U(1,2) + 0.5U(4,5)$ and $\frac{\sigma^2}{n} \sim U(0.1,1)$. This is a case where μ and σ^2 are independent and have a bimodal distribution. As in Example 11, the $N\Gamma^{-1}$ method outperforms all other methods with respect to estimating μ. However, presumably because of the distribution of σ^2 is unimodal, SURE-based methods do better in terms of estimating σ^2.*

Example 13. *The distribution of (μ, σ^2) is such that $\mu \sim N(3, 1^2)$ and $\sigma^2|\mu \sim U(max(\mu - 1, 0.1), max(\mu + 1, 1))$. Here, μ and σ^2 are dependent, which is a case where SURE-based methods do not perform well. The $N\Gamma^{-1}$ method outperforms the other methods in terms of $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ and in terms of $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ for larger q.*

Example 14. *The distribution of (μ, σ^2) is such that $\mu \sim N(3, 1^2)$ and $\sigma^2|\mu \sim max(N(\frac{|\mu|}{3}, (\frac{|\mu|}{3} + 1)^2), 0.1)$. Again, since μ and σ^2 are dependent, the SURE-based methods do not perform well. The group-linear algorithms lose efficiency as σ_i^2 is replaced by $S_{i\cdot}^2$, and the $N\Gamma^{-1}$ method outperforms all other methods in terms of both $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ and $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$.*

2.5 Real Data Example when Variance Matrix is Known

In this section, we consider a baseball data example as a test case for our mixture model method. This data set has been used in the articles of Brown [2008], Xie et al. [2012], Jing et al. [2016], and Weinstein et al. [2018]. The data consist of the entire season batting records for all major league baseball players in 2005 season. The goal is to estimate batting averages of individual players in the second half of the season by observing only the first half averages. Following the other articles, only players with at least 11 at-bats in the first half of the season were considered in the estimation process and only players with at least 11 at-bats in each of the two halves of the season were considered in the validation process.

Let H_{ij} denote the number of hits and N_{ij} the number of at-bats for player i in period j. The subscript j indicates either the first or second half of the season. The quantity p_i

Table 2.1: Averages of $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ over all $q = 20, 60, \ldots, 500$ in model (1.2) for Examples 1-8 of Section (2.4.1). For a given q, $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ is an average over 1000 replications.

Different Methods	Example 1	Example 2	Example 3	Example 4	Example 5	Example 6	Example 7	Example 8
Sample Statistics	0.5504	0.5496	0.5506	0.1248	0.3008	0.5502	0.5506	0.8976
EBMLE.XKB	0.3410	0.0762	0.0833	0.0071	0.2524	0.0814	0.4311	0.8448
EBMOM.XKB	0.3412	0.0832	0.0906	0.0086	0.2467	0.0822	0.4313	0.8423
JS.XKB	0.3675	0.0837	0.0885	0.0075	0.2616	0.085	0.4523	0.8563
Oracle.XKB	0.3328	0.0691	0.0535	0.0051	0.1936	0.0535	0.4258	0.7602
SURE.G.XKB	0.3424	0.0792	0.0645	0.0072	0.2365	0.0613	0.4327	0.8393
SURE.M.XKB	0.3433	0.0795	0.0639	0.0072	0.1988	0.0608	0.4334	0.7811
SURE.SG.XKB	0.3526	0.086	0.0699	0.0088	0.3068	0.0621	0.4561	0.8824
SURE.SM.XKB	0.3557	0.0877	0.0698	0.0091	0.1877	0.0628	0.4569	0.6829
GL.WMBZ	0.3512	0.098	0.0373	0.0141	0.1578	0.0306	0.4387	0.7401
GL.SURE.WMBZ	0.3534	0.0974	0.0473	0.0127	0.1578	0.0368	0.4415	0.7249
DGL.WMBZ	0.3714	0.1155	0.1044	0.0158	0.2496	0.0937	0.4523	0.8525
$N\Gamma^{-1}$mixture	0.3471	0.0894	0.0548	0.0102	0.1560	0.0532	0.2787	0.2639

Table 2.2: Averages of $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ over all $q = 20, 60, \ldots, 500$ in model (1.1) for Examples 9-14 of Section (2.4.2). For a given q, $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ is an average over 1000 replications.

Different Methods	Example 9	Example 10	Example 11	Example 12	Example 13	Example 14
Sample Statistics	0.1247	0.7484	0.1376	0.5520	0.7525	0.3570
EBMLE.XKB	0.1217	0.6369	0.1795	0.4681	0.4925	0.2432
EBMOM.XKB	0.1217	0.6381	0.1710	0.4690	0.4881	0.2430
JS.XKB	0.1222	0.6637	0.1328	0.5023	0.5997	0.3416
Oracle.XKB	0.1214	0.6350	0.1362	0.4670	0.4491	0.2342
SURE.G.XKB	0.1223	0.6479	0.1373	0.4765	0.4704	0.2428
SURE.M.XKB	0.1224	0.6483	0.1371	0.4769	0.4513	0.2384
SURE.SG.XKB	0.1249	0.6927	0.1343	0.5215	0.5461	0.2662
SURE.SM.XKB	0.1252	0.6954	0.1343	0.5228	0.5289	0.2638
GL.WMBZ	0.1216	0.6644	0.1317	0.4882	0.4958	0.2544
GL.SURE.WMBZ	0.1220	0.6720	0.1310	0.4965	0.5020	0.2589
DGL.WMBZ	0.1200	0.6045	0.1312	0.4538	0.4430	0.2679
SURE.M.Double	0.1199	0.5995	0.1319	0.4493	0.4340	0.2649
$N\Gamma^{-1}$ mixture	0.1198	0.5995	0.0911	0.2849	0.4176	0.2333

Table 2.3: Averages of $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ over all $q = 20, 60, \ldots, 500$ in model 1.1 for Examples 9-14 of Section (2.4.2). For a given q, $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ is an average over 1000 replications.

Different Methods	Example 9	Example 10	Example 11	Example 12	Example 13	Example 14
Sample Statistics	0.2206	6.6980	0.3836	3.9425	1.1918	2.9284
SURE.M.Double	0.0626	0.9160	0.1548	0.8692	0.2873	1.3108
$N\Gamma^{-1}$	0.0689	1.0065	0.1283	0.9630	0.3072	1.0025

Figure 2.2: $\widehat{MSE}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ vs. dimension q of normal vector for Examples 9-14 of Section 2.4.2. The dimension sizes are $q = 20, 60, \ldots, 500$ and results are based on 1000 replications at each q.

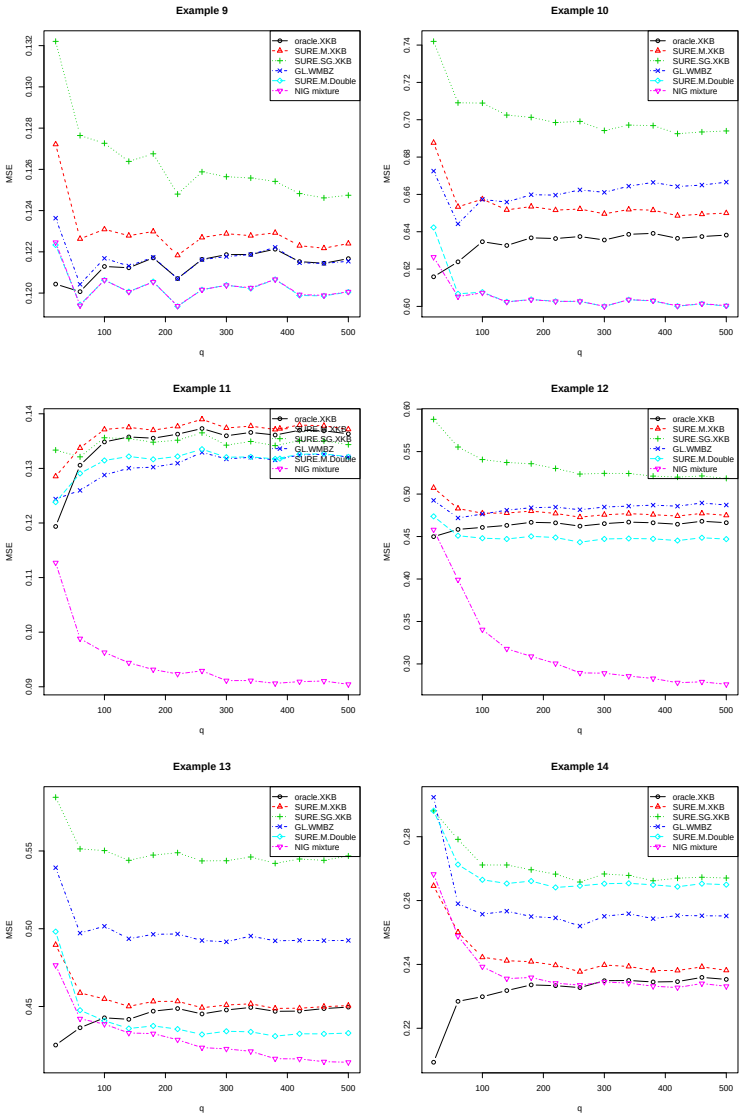
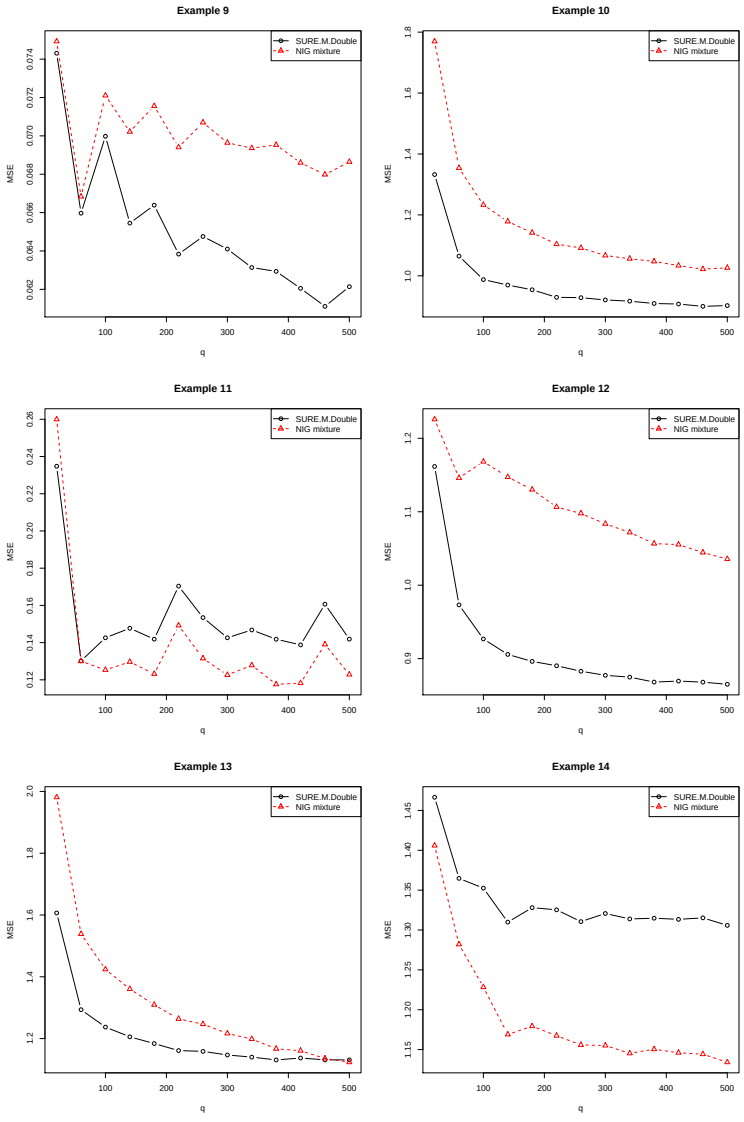


Figure 2.3: $\widehat{MSE}(\widehat{\boldsymbol{D}}, \boldsymbol{D})$ vs. dimension q of normal vector for Examples 9-14 of Section 2.4.2. The dimension sizes are $q = 20, 60, \ldots, 500$ and results are based on 1000 replications at each q.



denotes the probability of a hit for player i. Then we assume that

$$H_{ij} \sim \text{bin}(N_{ij}, p_i), \quad \text{for} \quad j = 1, 2, \quad i = 1, \ldots, q.$$

Without doing any variance-stabilizing transformation, Jing et al. [2016] worked with the sample proportion $X_{i1} = H_{i1}/N_{i1}$ and the estimated variance, $S_{i1}^2 = (X_{i1}(1 - X_{i1}))/N_{i1}$, of X_{i1}. However, this contradicts their initial assumption that X_{i1} and S_{i1}^2 are independently distributed. Also, without the transformation there is no reason to believe that X_{i1} is normally distributed and S_{i1}^2 follows a chi-square distribution. So, we will follow the transformation of Brown [2008], which was also used in Xie et al. [2012] and Weinstein et al. [2018], and define

$$X_{ij} = \arcsin\sqrt{\frac{H_{ij} + 0.25}{N_{ij} + 0.5}},$$

resulting in

$$X_{ij} \dot{\sim} N(\mu_i, \sigma_{ij}^2), \quad \mu_i = \arcsin(p_i), \quad \sigma_{ij}^2 = (4N_{ij})^{-1}.$$

The measure of error that was used in all these papers, denoted TSE, is used to compare different methods:

$$TSE(\hat{\boldsymbol{\mu}}) = \frac{\sum_i (X_{i2} - \hat{\mu}_i)^2 - \sum_i (4N_{i2})^{-1}}{\sum_i (X_{i2} - X_{i1})^2 - \sum_i (4N_{i2})^{-1}}.$$

The transformed data are consistent with model (1.2) as all σ_i^2 are known. The MCMC algorithm described in Section 2.5 is modified here by simply removing the step of updating σ_i^2. Table 2.4 is the table from Weinstein et al. [2018] with our method added in the bottom row.

Table 2.4: Average Prediction error for transformed batting averages. $TSE(\hat{\boldsymbol{\mu}})$ was computed for the entire data set, and separately for pitchers and non-pitchers from Weinstein et al. [2018].

Different	Data sets		
Methods	All	Pitchers	Non-pitchers
Naive	1	1	1
Grand mean	0.852	0.127	0.378
Nonparametric EB	0.508	0.212	0.372
Binomial mixture	0.588	0.156	0.314
Weighted Least Squares	1.07	0.127	0.468
Weighted nonparametric MLE	0.306	0.173	0.326
Weighted Least Squares (AB)	0.537	0.087	0.29
Weighted nonparametric MLE (AB)	0.301	0.141	0.261
JS.XKB	0.535	0.165	0.348
SURE.M.XKB	0.421	0.123	0.289
SURE.SG.XKB	0.408	0.091	0.261
GL.WMBZ	0.302	0.178	0.325
DGL.WMBZ	0.288	0.168	0.349
$N\Gamma^{-1}$ mixture	0.361	0.161	0.292

The naive estimator simply uses X_{i1} to predict X_{i2} and has TSE equal to 1. The grand mean uses the average of all X_{i1} to predict any X_{i2}. The nonparametric EB method of Brown and Greenshtein [2009], the binomial mixture of Muralidharan [2010], the weighted least squares estimator, the weighted least squares estimator (AB) (with number of at-bats as covariate), the weighted nonparametric MLE and the weighted nonparametric MLE (AB) (with number of at-bats as covariate) of Jiang et al. [2009] are also included in Table 2.4.

Weinstein et al. [2018] presented an analysis under permutations, where each permutation is the order in which successful hits appear throughout the entire season. For each player they draw the number of hits in N_{i1} at bats from a hypergeometric distribution, $HG(N_{i1} + N_{i2}, H_{i1} + H_{i2}, N_{i1})$. We compare our method with several other methods with respect to 1000 different permutations of the baseball data and average TSE.

As discussed in Weinstein et al. [2018], group linear algorithms tend to perform well compared to SURE-based methods as μ_i and σ_{i1}^2 are not independent, owing to the fact that

players with higher batting averages tend to play more. Also, non-pitchers tend to have higher batting averages than pitchers, so it is possible that the underlying density of μ is bimodal. This may be the reason that empirical Bayes estimators that assume a normal-normal model tend to perform poorly. Group linear estimates outperform the other methods because they can accommodate these features exhibited by the baseball data. SURE-based methods work well when we analyze the pitchers and non-pitchers separately. Table 2.4 shows that, in the combined data, the $N\Gamma^{-1}$ method does not perform as well as group linear algorithms, but it performs better than SURE-based methods. However, when the pitchers and non-pitchers are considered separately, $N\Gamma^{-1}$ performs better than the group linear algorithms. In both the original data and the permuted data, $N\Gamma^{-1}$ performs better than the group linear algorithms for both pitchers and non-pitchers. When pitchers and non-pitchers are combined, group linear methods outperform all other methods in both the original and permuted data. This is reasonable as the association between μ and σ^2 is weaker when the data are separated into smaller groups, and group linear algorithms work well in the presence of strong association. In contrast, the $N\Gamma^{-1}$ method works reasonably well μ and σ^2 are either strongly or weakly dependent.

Table 2.5: Average Prediction error for 1000 permutations of transformed batting averages data. Average $TSE(\hat{\boldsymbol{\mu}})$ was computed for the entire data set, and separately for pitchers and non-pitchers.

Different	Data sets		
Methods	All	Pitchers	Non-pitchers
Grand mean	0.9222	0.3127	0.2951
James-Stein	0.5465	0.2490	0.2304
SURE.M.XKB	0.4852	0.2227	0.2602
SURE.SG.XKB	0.4693	0.1759	0.2148
GL.WMBZ(bins = $q^{1/3}$)	0.2798	0.2438	0.1731
GL.SURE.WMBZ	0.3032	0.2838	0.1949
DGL.WMBZ	0.4751	0.2193	0.2250
$N\Gamma^{-1}$ mixture	0.3535	0.2377	0.1698

2.6 Real Data Example when Variance Matrix is Unknown

In this section, we will apply the $N\Gamma^{-1}$ method and other estimators to the prostate data from the book of Efron [2012]. The data can be downloaded from the book website:

https://statweb.stanford.edu/ ckirby/brad/LSI/datasets-and-programs/datasets.html.

The prostate data consist of gene expression levels for $q = 6033$ genes obtained from 102 men, 50 normal control and 52 prostate cancer patients. We only use the control data, which means that we have a 6033×50 matrix. Here X_{ij} denotes the expression level for gene i of patient j, $i = 1, \ldots, 6033$, $j = 1, \ldots, 50$. Since 50 is a relatively large number, we will assume that the control group constitutes the population of interest, in which case

$$\mu_i = \frac{1}{50}\sum_{j=1}^{50} X_{ij} \quad \text{and} \quad \sigma_i^2 = \frac{1}{50}\sum_{j=1}^{50}(X_{ij} - \mu_i)^2, \quad i = 1, \ldots, 6033.$$

As a test of the various methods, we randomly select three subjects from the control group and use their data to estimate μ_i and σ_i^2.

To better understand the nature of the data we provide the scatterplots in Figures 2.4-2.5. We also compared our method with the sample means and variances from three columns.

To compare different methods we randomly chose 500 rows and 3 columns, computed estimates of means and variances using the various methods, and replicated this process 100 times. Average squared error for each method was computed as in our simulation study. Table 2.6 shows that, except for the SURE-based Double shrinkage estimators, all methods were outperformed by $N\Gamma^{-1}$. Figure 2.6 shows that the densities of μ_i and σ_i^2 are well-approximated by normal and inverse gamma densities, respectively. When we force the mixture of normal-inverse gammas to select only one component, then this method performs comparably to SURE.M.Double for estimating both $\boldsymbol{\mu}$ and $\boldsymbol{D}$. For the other algorithms, replacing the unknown σ_i^2 with S_i^2 result in a loss in accuracy of those methods.

Figure 2.4: Scatterplots for prostate data. The upper left plot is $S_{i\cdot}^2$ vs. $\bar{X}_{i\cdot}$ for columns 6, 30 and 31 of the data matrix, the upper right plot is σ_i^2 vs. μ_i and the lower left plot is $\hat{\sigma}_{i,DPMM}^2$ vs. $\hat{\mu}_i^{DPMM}$ based on columns 6, 30 and 31.

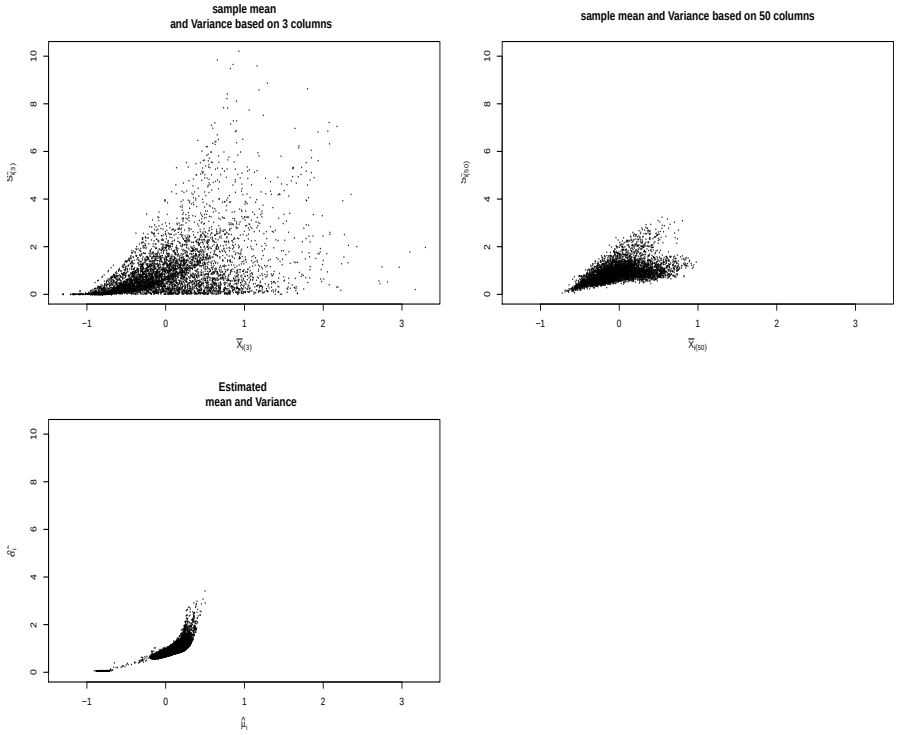


Figure 2.5: Scatterplots for prostate data based 3 columns 6, 30 and 31 of the data matrix. The upper left plot is $\bar{X}_{i\cdot}$ vs. μ_i, the upper right plot is $\hat{\mu}_i^{DPMM}$ vs. μ_i based on columns 6, 30 and 31. The lower left plot is $\bar{S}_{i\cdot}^2$ vs. σ_i^2, the lower right plot is $\hat{\sigma}_{i,DPMM}^2$ vs. σ_i^2 based on columns 6, 30 and 31.

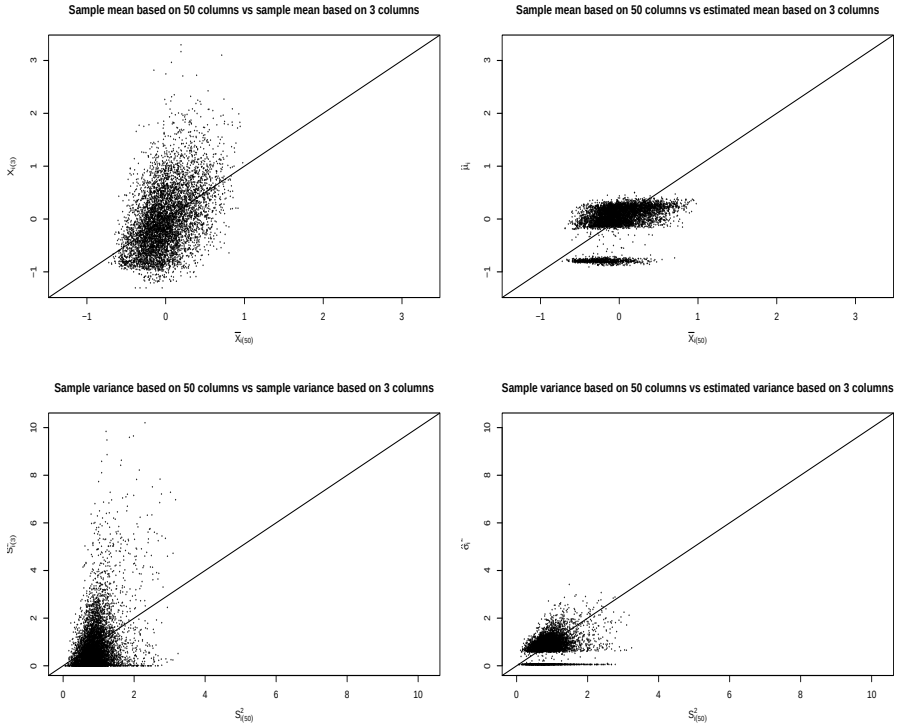
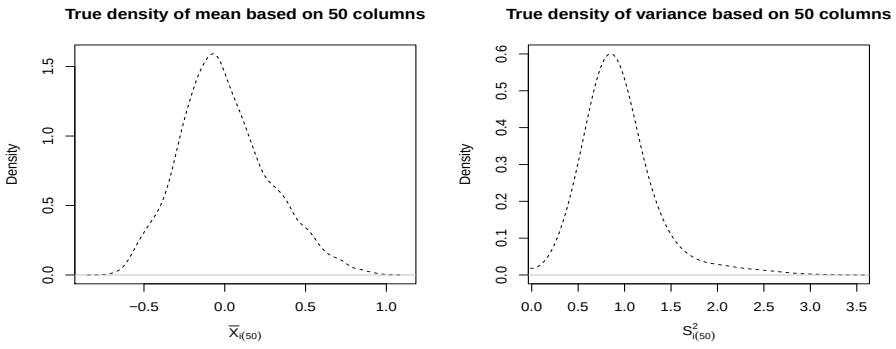


Table 2.6: Estimated average squared loss for $\boldsymbol{\mu}$ and $\boldsymbol{D}$ for different estimation methods from prostate-control data. Each table value is an average over 100 replications. Each replication consists of 500 randomly chosen rows and 3 randomly chosen columns from the original 6033×50 data matrix.

Different	Different measures	
Methods	Error in estimating μ_i	Error in estimating σ_i^2
Sample Statistics	0.2919	1.1695
EBMLE.XKB	0.1486	-
EBMOM.XKB	0.1446	-
JS.XKB	0.2787	-
Oracle.XKB	0.1108	-
SURE.G.XKB	0.1071	-
SURE.M.XKB	0.1175	-
SURE.SG.XKB	0.1445	-
SURE.SM.XKB	0.1682	-
GL.WMBZ	0.1694	-
GL.SURE.WMBZ	0.1802	-
DGL.WMBZ	0.0690	-
SURE.M.Double	0.0644	0.1458
$N\Gamma^{-1}$ mixture	0.1081	0.2284
$N\Gamma^{-1}$ one component	0.0683	0.1653

Figure 2.6: Marginal kernel density estimates computed from μ_i and σ_i^2 based on all 50 columns of the data matrix.



3. LOCATION-SCALE DENSITY ESTIMATION USING MIXTURE

As mentioned in Section 1.2, the main focus of this chapter is a semiparametric estimation of the joint density of μ and σ^2 in the LSRE model. In this section, we discuss a method of estimating the density f_{μ,σ^2} when f_ϵ is standard normal.

3.1 Density Estimation for the LSRE Model with Normal Error Density

To estimate the distribution of the bivariate density f_{μ,σ^2} nonparametrically, it is reasonable to use a mixture of bivariate densities, which underlies most mainstream approaches of density estimation, including kernel techniques (Silverman [1986]), nonparametric maximum likelihood (Lindsay et al. [1983]), and Bayesian approaches using mixtures induced by a Dirichlet process (Ferguson [1983] and Escobar and West [1995]). Here, we assume f_{μ,σ^2} is a mixture of normal-inverse gamma densities, and we are trying to estimate the parameters of the mixture distribution.

3.1.1 Identifiability of the Joint Distribution of Location-Scale Parameters

The seminal paper of Reiersøl [1950] showed that in the LRE model, both f_μ and f_ϵ are identifiable from the joint distribution of (X_{i1}, X_{i2}). Hart and Cañette [2011] showed that, in the LSRE model, under some regularity conditions, if $n \geq 4$ then the $\log(\sigma_i)$, ϵ_{ij}, and μ_i distributions are all identifiable from the joint distribution of $(X_{i1}, X_{i2}, X_{i3}, X_{i4})$. Beran and Millar [1994] showed that f_{μ,σ^2} is identifiable from the joint distribution of (X, ϵ). In our case ϵ is not observed, and hence we initially assume that the density of ϵ is normal. Teicher [1960] showed that if f_ϵ is normal, f_μ and f_{σ^2} are not both identifiable from the density of X_{i1}.

Suppressing dependency on i, when all σ^2s are known and f_ϵ is standard normal, in model (1.6) with $n = 1$, then f_{μ,σ^2} is identifiable from the joint density of (X_1, σ^2), f_{X_1,σ^2}. We can write $X_1 = \mu + e_1$, where, $e_1 = \sigma\epsilon_1$. From f_{X_1,σ^2}, we know the marginal density f_{σ^2} and the conditional density $f_{X_1|\sigma^2}$. Conditional on σ^2, μ and e are independent and they

add up to X_1. The characteristic function of $X_1|\sigma^2$, $\phi_{X_1|\sigma^2}(t) = \int_{\mathbb{R}} e^{itx} f_{X|\sigma^2}(x)dx$, equals $\phi_{\mu|\sigma^2}(t)\phi_{e|\sigma^2}(t)$, for all t. As the characteristic function of $e|\sigma^2$, $\phi_{e|\sigma^2}$, is known and never vanishes, we can uniquely identify $\phi_{\mu|\sigma^2}$ from $\phi_{X_1|\sigma^2}$. From uniqueness of the characteristic function, $f_{\mu|\sigma^2}$ identifiable. So, we can uniquely identify f_{μ,σ^2} from f_{X_1,σ^2}.

When σ^2s are unknown, then we need $n = 2$ to identify f_{μ,σ^2} from model (1.6). Let (X_1, X_2) follow model (1.6) with $n = 2$. Let F_{μ,σ^2} be the bivariate cumulative distribution function (c.d.f.) of random variable (μ, σ^2). When f_ϵ is standard normal, the joint density of (X_1, X_2), f_{X_1,X_2}, is

$$\begin{aligned} f_{X_1,X_2}(x,y) &= \int_{\mathbb{R}^+}\int_{\mathbb{R}} \frac{1}{v} f_\epsilon\left(\frac{x-m}{\sqrt{v}}\right) f_\epsilon\left(\frac{y-m}{\sqrt{v}}\right) dF_{\mu,\sigma^2}(m,v) \\ &= \int_{\mathbb{R}^+}\int_{\mathbb{R}} \frac{1}{2\pi v} e^{-\frac{1}{2v}\left((x-m)^2+(y-m)^2\right)} dF_{\mu,\sigma^2}(m,v). \end{aligned} \tag{3.1}$$

Let a finite probability measure be a discrete probability distribution with a finite number of atoms. It has been shown by Teicher [1960] that if F_{μ,σ^2} is restricted to the class of finite probability measures, F_{μ,σ^2} is identifiable from the marginal density of f_{X_1}. If F_{μ,σ^2} is unrestricted, then it is not identifiable from the marginal density of f_{X_1}. It will now be shown that if we restrict F_{μ,σ^2} to the class of absolutely continuous probability measures, then the solution F_{μ,σ^2} of (3.1) is unique.

Theorem 1. *If f_ϵ is a standard normal density and f_{σ^2} is a bounded and continuous density on $(0,\infty)$, then f_{μ,σ^2} is identifiable from f_{X_1,X_2}.*

Proof 1. *The characteristic function of the joint density for* (X_1, X_2) *is*

$$\begin{aligned}
\phi_{X_1,X_2}(t_1,t_2) &= \int_{\mathbb{R}^+}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{it_1X_1+it_2X_2} f_\epsilon(e_1)f_\epsilon(e_2)de_1de_2dF_{\mu,\sigma^2}(m,v) \\
&= \int_{\mathbb{R}^+}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{im(t_1+t_2)+i\sqrt{v}(t_1e_1+t_2e_2)} f_\epsilon(e_1)f_\epsilon(e_2)de_1de_2f_{\mu,\sigma^2}(m,v)dmdv \\
&= E(e^{i\mu(t_1+t_2)+i\sigma(t_1\epsilon_1+t_2\epsilon_2)}) = E_{\sigma^2}(E(e^{i\mu(t_1+t_2)}|\sigma^2)E(e^{i\sigma(t_1\epsilon_1+t_2\epsilon_2)}|\sigma^2)) \\
&= \int_{\mathbb{R}^+}\int_{\mathbb{R}} e^{-\frac{1}{2}(t_1^2+t_2^2)v} e^{i(t_1+t_2)m} f_{\mu|\sigma^2=v}(m)dm f_{\sigma^2}(v)dv \\
&= \int_{\mathbb{R}^+} e^{-\frac{1}{2}(t_1^2+t_2^2)v} \left(\int_{\mathbb{R}} e^{i(t_1+t_2)m} f_{\mu,\sigma^2}(m,v)dm\right) dv.
\end{aligned}$$

Without loss of generality, we may assume that $t_1 < t_2$ *as* $\phi_{X_1,X_2}(t_1,t_2) = \phi_{X_1,X_2}(t_2,t_1)$. *We may define* u_1 *and* u_2 *as* t_1+t_2 *and* $(t_1^2+t_2^2)/2 > 0$, *respectively and rewrite* t_1 *and* t_2 *as* $(u_1 - \sqrt{4u_2 - {u_1}^2})/2$ *and* $(u_1 + \sqrt{4u_2 - {u_1}^2})/2$, *respectively. Then for any* ${u_2} > {u_1}^2/4$, *the function* $\phi_{X_1,X_2}(t_1,t_2)$ *can be written as a function of* u_1 *and* u_2

$$\phi_{X_1,X_2}(u_1,u_2) = \int_{\mathbb{R}^+} e^{-u_2v}\left(\int_{\mathbb{R}} e^{iu_1m} f_{\mu,\sigma^2}(m,v)dm\right)dv = \mathcal{L}(f_1(v|u_1))(u_2),$$

where $f_1(v|u_1) = \int_{\mathbb{R}} e^{iu_1m} f_{\mu,\sigma^2}(m,v)dm = f_{\sigma^2}(v)\phi_{\mu|\sigma^2}(u_1|v)$ *and* $\phi_{\mu|\sigma^2}$ *denotes the characteristic function of the random variable* μ *conditional on* σ^2. *The characteristic function* ϕ_{X_1,X_2} *is the (unilateral) Laplace transform of the inside integral* $f_1(v|u_1)$. *Suppose that the joint densities* f_{μ,σ^2} *and* g_{μ,σ^2} *have the same characteristic function* ϕ_{X_1,X_2}. *We will now prove that* $f_{\mu,\sigma^2} \equiv g_{\mu,\sigma^2}$. *Let us define* $g_1(v|u_1) = \int_{\mathbb{R}} e^{iu_1m} g_{\mu,\sigma^2}(m,v)dm$ *and from the uniqueness of the Laplace transformation we have* $f_1(v|u_1) = g_1(v|u_1)$ *if* f_{σ^2} *is bounded and continuous.*

For every $u_1 < 2\sqrt{u_2}$,

$$\int_{\mathbb{R}} e^{iu_1 m} f_{\mu,\sigma^2}(m,v)dm = \int_{\mathbb{R}} e^{iu_1 m} g_{\mu,\sigma^2}(m,v)dm, \quad \forall v$$

$$\text{or,} \quad \int_{\mathbb{R}^+} e^{iw_1 v} \int_{\mathbb{R}} e^{iu_1 m} f_{\mu,\sigma^2}(m,v)dmdv = \int_{\mathbb{R}^+} e^{iw_1 v} \int_{\mathbb{R}} e^{iu_1 m} g_{\mu,\sigma^2}(m,v)dmdv, \quad \forall (w_1,u_1)$$

$$\text{or,} \quad \phi_f(u_1,w_1) = \phi_g(u_1,w_1).$$

From the uniqueness of the characteristic function it follows that $f_{\mu,\sigma^2} = g_{\mu,\sigma^2}$.

The fact that the characteristic function of a normal distribution never vanishes makes it easier to prove that f_{μ,σ^2} is identifiable. The necessary and sufficient conditions that are needed on f_ϵ for the identifiability of F_{μ,σ^2} in (3.1) is still an open question.

3.2 Modeling the Joint Distribution of Location-Scale by a Bivariate Mixture

In Section 2.1 and 2.2, a MCMC algorithm is proposed to estimate location-scale parameters. Furthermore, at every MCMC iteration we may obtain estimates of $(\boldsymbol{\Theta}_r, \boldsymbol{\pi}_r)$ for $r = 1, \ldots, k$, from which we can calculate values of the mixture density over a grid. Averaging density values over all iterations leads to an estimate of f_{μ,σ^2}.

3.3 Simulation Study

The mean integrated squared error (MISE) measures the overall accuracy of estimating f_{μ,σ^2} by $\hat{f}_{\mu,\sigma^2}$. We can estimate MISE using Monte Carlo methods and B simulated datasets. Letting $\hat{f}^b_{\mu,\sigma^2}$ denote the density estimate from the b^{th} set of simulated data, the MISE and its estimate are defined as

$$\begin{aligned} MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2}) &= \int_{\mathbb{R}^+}\int_{\mathbb{R}} E\Big(f_{\mu,\sigma^2}(m,v) - \hat{f}_{\mu,\sigma^2}(m,v)\Big)^2 dmdv \\ &\approx \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(f_{\mu,\sigma^2}(m_i,v_j) - \hat{f}^b_{\mu,\sigma^2}(m_i,v_j)\Big)^2 |m_i - m_{i-1}||v_j - v_{j-1}|, \end{aligned}$$

where $\{m_i, v_j\}_{(i,j)=(0,0)}^{(N,N)}$ are grid points that are evenly spaced in each direction over the support of f_{μ,σ^2}. We will use MISE to measure the performance of different methods for several choices of f_{μ,σ^2}. Similarly, using the same grid points, we may compute the MISE when estimating the marginal densities, f_μ and f_{σ^2}.

3.3.1 Comparing the Density Estimate with Other Density Estimators

To our knowledge there exist no competitors in the literature for our method of estimating the density of f_{μ,σ^2}. As mentioned previously, Staudenmayer et al. [2008] and Sarkar et al. [2014] took an alternative Bayesian approach in which μ is a random effect, the variance is a function of μ and the scaled error density is standard normal. Sarkar et al. [2014] also relaxed the normality assumption on the scaled error density and estimated the scaled error density with a mixture of normal densities. The code for the methodology of Staudenmayer et al. [2008] was not available, however, we were able to compare our method with that of Sarkar et al. [2014], which we refer to here as DPMM.SMSPC. The model of DPMM.SMSPC entails that the distribution of (μ, σ^2) is singular, and hence implicitly provides an estimate of this singular distribution.

A possible means of estimating f_{μ,σ^2} would be to estimate (μ_i, σ_i^2), $i = 1, \ldots, q$, and to then compute a kernel density estimator (KDE) from the estimates as if they were the true values of (μ_i, σ_i^2). We will call such estimates "plug-in KDE" There are a few articles, including Xie et al. [2012] and Weinstein et al. [2018], which address the problem of estimating μ_i when σ_i^2 is known. Jing et al. [2016] further extended the work of Xie et al. [2012] and estimated both μ_i and σ_i^2. The estimators of Xie et al. [2012] defined by their expressions (7.1), (7.2), (7.3), (4.2), (5.1), (6.3), and (6.2) will be called EBMLE.XKB, EBMOM.XKB, JS.XKB, SURE.G.XKB, SURE.M.XKB, SURE.SG.XKB, and SURE.SM.XKB, respectively. The methods which were referred to as SURE.M.Double can be found in expressions (11-12) of Jing et al. [2016]. Weinstein et al. [2018] developed group-linear and dynamic group-linear algorithms, which are referred to here as GL.WMBZ and DGL.WMBZ, respectively. We also consider Oracle.XKB, which, although not an estimate as described in section 7 of Xie et al.

[2012], provides a sensible lower bound on a mean squared risk estimator of given parametric form.

We simulated data from model (1.6) for different choices of f_{μ,σ^2}. For each (μ_i, σ_i^2) pair there are $n = 4$ replications. Importantly, only the observations X_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, n$, are used to define estimates of f_{μ,σ^2}. Let $\boldsymbol{y} = (y_1, \ldots, y_k)^T$, then $\text{IQR}(\boldsymbol{y})$, $\min(\boldsymbol{y})$, and $\max(\boldsymbol{y})$ denote the interquartile range, minimum, and maximum of observations $y_1, \ldots, y_k$, respectively. Also, define $\overline{\boldsymbol{X}} = (\bar{X}_{1\cdot}, \ldots, \bar{X}_{q\cdot})^T$ and $\boldsymbol{S}^2 = (S_{1\cdot}^2, \ldots, S_{q\cdot}^2)^T$. Then for each data set, we divided the range $[\min(\overline{\boldsymbol{X}}) - \text{IQR}(\overline{\boldsymbol{X}}), \max(\overline{\boldsymbol{X}}) + \text{IQR}(\overline{\boldsymbol{X}})] \times [\max(\min(\boldsymbol{S}^2) - \text{IQR}(\boldsymbol{S}^2), 0.001), \max(\boldsymbol{S}^2) + \text{IQR}(\boldsymbol{S}^2)]$ into 100×100 equally spaced grid points and calculated, for each of the different methods, the approximate MISEs for estimates of both joint and marginal distributions. The ensuing tables compare the MISEs of various methods. Our method based on the mixture of normal-inverse gamma distributions is denoted $N\Gamma^{-1}$. This method produces estimates of all (μ_i, σ_i^2) pairs, and hence we may compute a plug-in estimate using these estimates. This method is referred to as $N\Gamma^{-1}$ KDE. We also consider plug-in estimates based on the methods of Xie et al. [2012], Jing et al. [2016], and Weinstein et al. [2018]. Since Xie et al. [2012] and Weinstein et al. [2018] assumed that $\sigma_1^2, \ldots, \sigma_q^2$ are known, we use $S_{1\cdot}^2, \ldots, S_{q\cdot}^2$ in our plug-in estimates for these methods. As the support of σ^2 is $(0, \infty)$, the kernel density estimator exhibits boundary bias near $\sigma^2 = 0$. To eliminate most of this bias, we reflect the data points around $\sigma^2 = 0$, and then compute the kernel density estimator using the resulting $2q$ observations and the default bandwidth in the R command `density` from `base R` and `kde2d` from library `MASS` for univariate and bivariate density estimation, respectively. We repeated the experiment B times for each q and plotted estimated MISEs for each q.

Example 15. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0, 3)$ and $\sigma^2 \sim IG(5, 2)$. Here and in all other examples, we take $\epsilon \sim N(0, 1)$. In this example, as the underlying true model is normal and inverse-gamma, the $N\Gamma^{-1}$ method performs better than the others. Figures 3.1 and 3.3 show that $N\Gamma^{-1}$ KDE outperforms the other plug-in methods*

when estimating the bivariate density and the marginal density of σ^2. When estimating the marginal density of μ, DPMM.SMSPC slightly outperforms the mixture of normal-inverse gamma densities since it is also based on a mixture of normals.

Example 16. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,3)$ and $\sigma^2 \sim G(9,3)$. This example is quite similar to Example 15. All methods work well in estimating the joint density f_{μ,σ^2}. As in Example 15, $N\Gamma^{-1}$ has lower MISE compared to the other methods. $N\Gamma^{-1}$ KDE performs similarly to SURE.M.Double as $N\Gamma^{-1}$ successfully indentifies only one component. When estimating the marginal density f_μ, both DPMM.SMSPC and $N\Gamma^{-1}$ perform better than the other methods.*

Example 17. *Here f_{μ,σ^2} is such that $(\mu,\sigma^2) \sim 0.95N\Gamma^{-1}(2,2,5,2) + 0.05N\Gamma^{-1}(10,4,3,3)$. As our method is based on a $N\Gamma^{-1}$ mixture, it performs better than the other methods for estimating all three densities, f_{μ,σ^2}, f_μ, and f_{σ^2}.*

Example 18. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim 0.5U(1,2) + 0.5U(4,5)$ and $\sigma^2/n \sim U(0.1,1)$, where $U(a,b)$ denotes the uniform distribution on the interval (a,b). In this example f_μ is bimodal, and $N\Gamma^{-1}$ and $N\Gamma^{-1}$ KDE estimate the densities better than the other methods. SURE based and group-linear methods shrink all observations towards the grand mean, resulting in poor KDEs compared to $N\Gamma^{-1}$ KDE. In contrast, the $N\Gamma^{-1}$ estimate of μ_i tends to shrink towards the mean of the component from which μ_i came, leading to better performance of the kernel estimate.*

Example 19. *The density f_{μ,σ^2} is such that $\mu \sim Inv-\chi^2_4$ and $log(\sigma^2)|\mu \sim N(log(\mu), 0.5^2)$. $Inv-{\chi_4}^2$ denotes an inverse-chi square distribution with four degrees of freedom. In this example, σ^2 is not independent of μ, and $N\Gamma^{-1}$ and $N\Gamma^{-1}$ KDE estimate the densities better than the other methods.*

Example 20. *The density f_{μ,σ^2} is such that $\mu \sim \Gamma(1,0.5)$ and $\sigma^2|\mu \sim \chi^2_\mu + 0.1$, where ${\chi_\mu}^2$ denotes a chi-square density with μ degrees of freedom. Here also σ^2 is not independent of μ, and $N\Gamma^{-1}$ and $N\Gamma^{-1}$ KDE perform better than the plug-in estimators when estimating*

f_μ. Since DPMM.SMSPC assumes that σ^2 is fixed by μ, and here the conditional variance of σ^2 increases with μ, a poor estimate of f_μ results from using the DPMM.SMSPC method.

Table 3.1: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$ and f_{μ,σ^2} defined by Examples 15-20 of Section 3.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 100, 200, \ldots, 1000$.

Different	MISE of f_{μ,σ^2}					
Methods	Example 15	Example 16	Example 17	Example 18	Example 19	Example 20
Sample Statistics	0.1438	0.0311	0.5639	0.1067	3.7866	0.2685
EBMLE.XKB	0.1424	0.0298	0.5581	0.1041	3.4922	0.2655
EBMOM.XKB	0.1424	0.0297	0.5587	0.1041	3.6907	0.2661
JS.XKB	0.1433	0.0306	0.5588	0.1055	3.5760	0.2665
Oracle.XKB	0.1424	0.0297	0.5624	0.1041	3.7854	0.2685
SURE.G.XKB	0.1423	0.0295	0.5625	0.1042	3.7656	0.2676
SURE.M.XKB	0.1423	0.0295	0.5623	0.1042	3.7856	0.2684
SURE.SG.XKB	0.1424	0.0298	0.5571	0.1046	3.6578	0.2661
SURE.SM.XKB	0.1425	0.0298	0.5576	0.1046	3.7274	0.2688
GL.WMBZ	0.1424	0.0297	0.5552	0.1044	3.4789	0.2666
GL.SURE.WMBZ	0.1424	0.0298	0.5533	0.1045	3.4745	0.2666
DGL.WMBZ	0.1424	0.0298	0.5556	0.1045	3.6647	0.2618
SURE.M.Double	0.0813	0.0115	0.4121	0.0946	3.9727	0.3374
$N\Gamma^{-1}$ mixture	0.0117	0.0046	0.032	0.0682	2.1160	0.3105
$N\Gamma^{-1}$ KDE	0.0789	0.0125	0.3162	0.0776	2.8536	0.6678

3.3.2 Analysis of Prostate Cancer Data

We will apply our method along with other estimators to the prostate data from the book of Efron [2012]. The data can be downloaded from the book website:

https://statweb.stanford.edu/ ckirby/brad/LSI/datasets-and-programs/datasets.html.

The prostate data consist of gene expression levels for $q = 6033$ genes obtained from 102 men, 50 normal control and 52 prostate cancer patients.

We use only the control data, which means that we have a 6033×50 matrix. Here, X_{ij} denotes the expression level for gene i on patient j, $i = 1, \ldots, 6033$, $j = 1, \ldots, 50$. Since 50 is a relatively large number, we will assume that the control group constitutes the population

Figure 3.1: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ vs. dimension q for Examples 15-20 of Section 3.3.1. Dimension size is $q = 100, 200, \ldots, 1000$ and number of replications is 100 for each q.

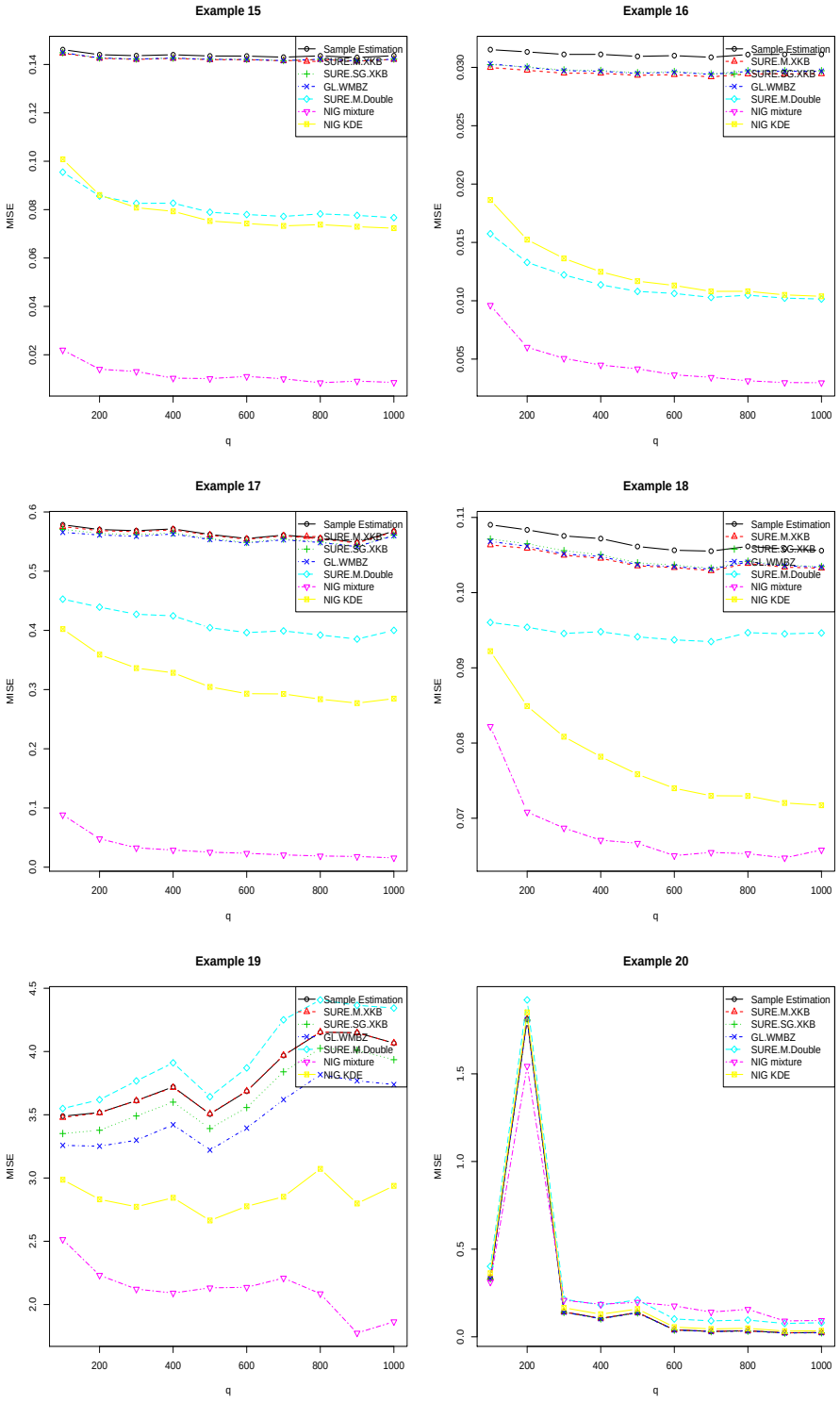


Table 3.2: Estimated $MISE(\hat{f}_\mu, f_\mu)$ averaged over values of q. The data were generated from (1.6) with $n = 4$ and f_{μ,σ^2} defined by Examples 15-20 of Section 3.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 100, 200, \ldots, 1000$.

Different	MISE of f_μ					
Methods	Example 15	Example 16	Example 17	Example 18	Example 19	Example 20
Sample Statistics	0.0016	0.0035	0.0238	0.237	0.3733	0.0170
EBMLE.XKB	0.0014	0.0022	0.0191	0.2325	0.2837	0.0176
EBMOM.XKB	0.0014	0.0024	0.0195	0.2337	0.3207	0.0172
JS.XKB	0.0015	0.0019	0.02	0.2283	0.2684	0.0157
Oracle.XKB	0.0014	0.0023	0.0227	0.2329	0.3876	0.0212
SURE.G.XKB	0.0014	0.0035	0.0226	0.2412	0.3626	0.0169
SURE.M.XKB	0.0014	0.0036	0.0227	0.2412	0.3943	0.0220
SURE.SG.XKB	0.0014	0.0043	0.0191	0.2492	0.3082	0.0177
SURE.SM.XKB	0.0014	0.0044	0.0195	0.2482	0.3633	0.0216
GL.WMBZ	0.0014	0.0031	0.0182	0.2396	0.2847	0.0197
GL.SURE.WMBZ	0.0014	0.0033	0.0173	0.2416	0.2966	0.0198
DGL.WMBZ	0.0014	0.0024	0.018	0.2291	0.2957	0.0220
SURE.M.Double	0.0014	0.0027	0.0185	0.2301	0.2925	0.0234
DPMM.SMSPC	0.0006	0.0016	0.0072	0.0881	0.2352	0.0414
$N\Gamma^{-1}$ mixture	0.0009	0.0008	0.0033	0.117	0.1653	0.0144
$N\Gamma^{-1}$ KDE	0.0014	0.0026	0.0196	0.1228	0.1479	0.0192

Table 3.3: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$ and f_{μ,σ^2} defined by Examples 15-20 of Section 3.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 100, 200, \ldots, 1000$.

Different	MISE of f_{σ^2}					
Methods	Example 15	Example 16	Example 17	Example 18	Example 19	Example 20
Sample Statistics	0.8664	0.1845	0.7887	0.1403	0.0493	0.0312
SURE.M.Double	0.4786	0.0695	0.4839	0.114	0.1263	0.2224
$N\Gamma^{-1}$ mixture	0.032	0.0166	0.0348	0.096	0.6236	0.1346
$N\Gamma^{-1}$ KDE	0.4826	0.0758	0.469	0.1009	0.282075	0.0906

Figure 3.2: Estimated $MISE(\hat{f}_\mu, f_\mu)$ vs. dimension q of normal vector for Examples 15-20 of Section 3.3.1. Dimension size is $q = 100, 200, \ldots, 1000$ and number of replications is 100 for each q.

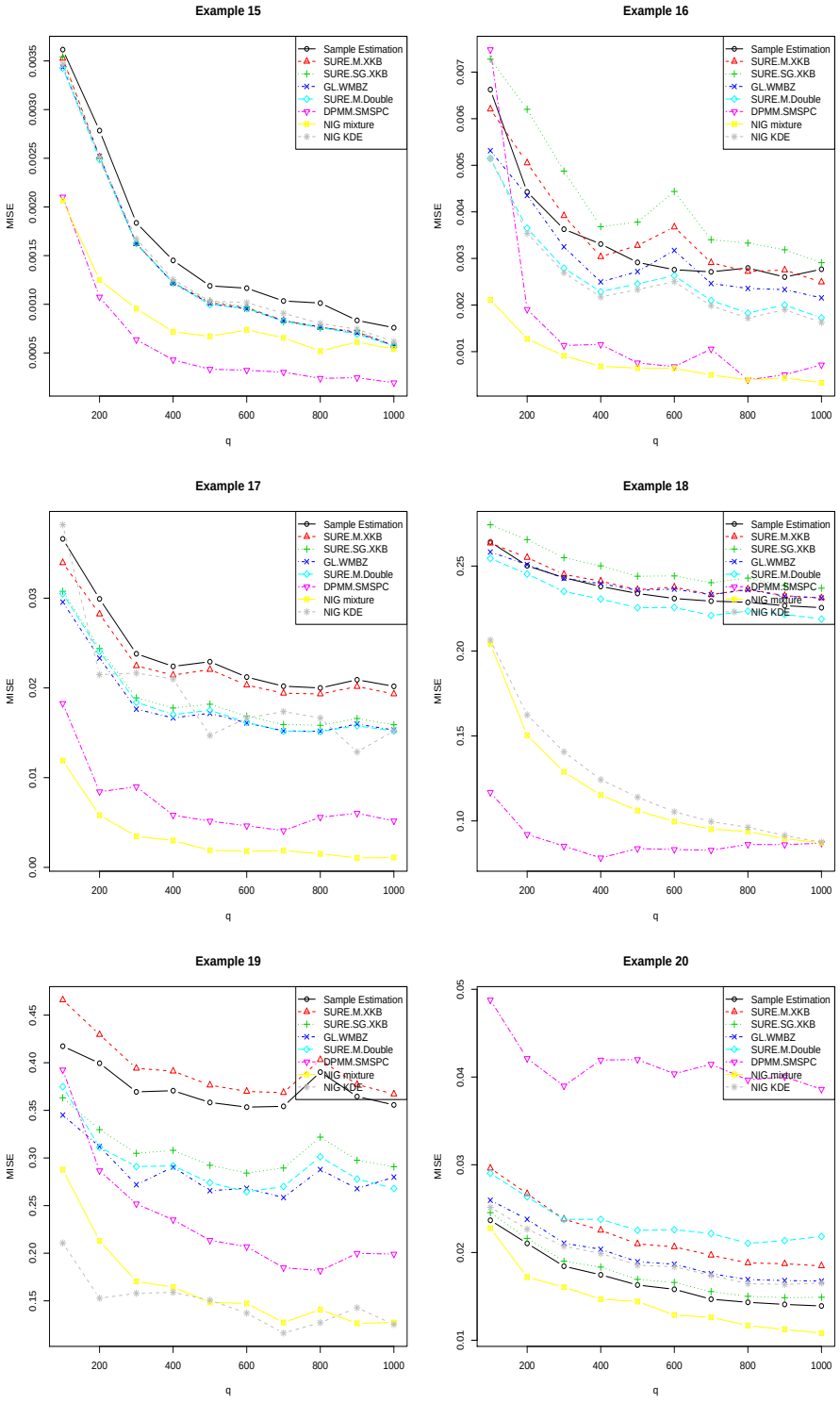
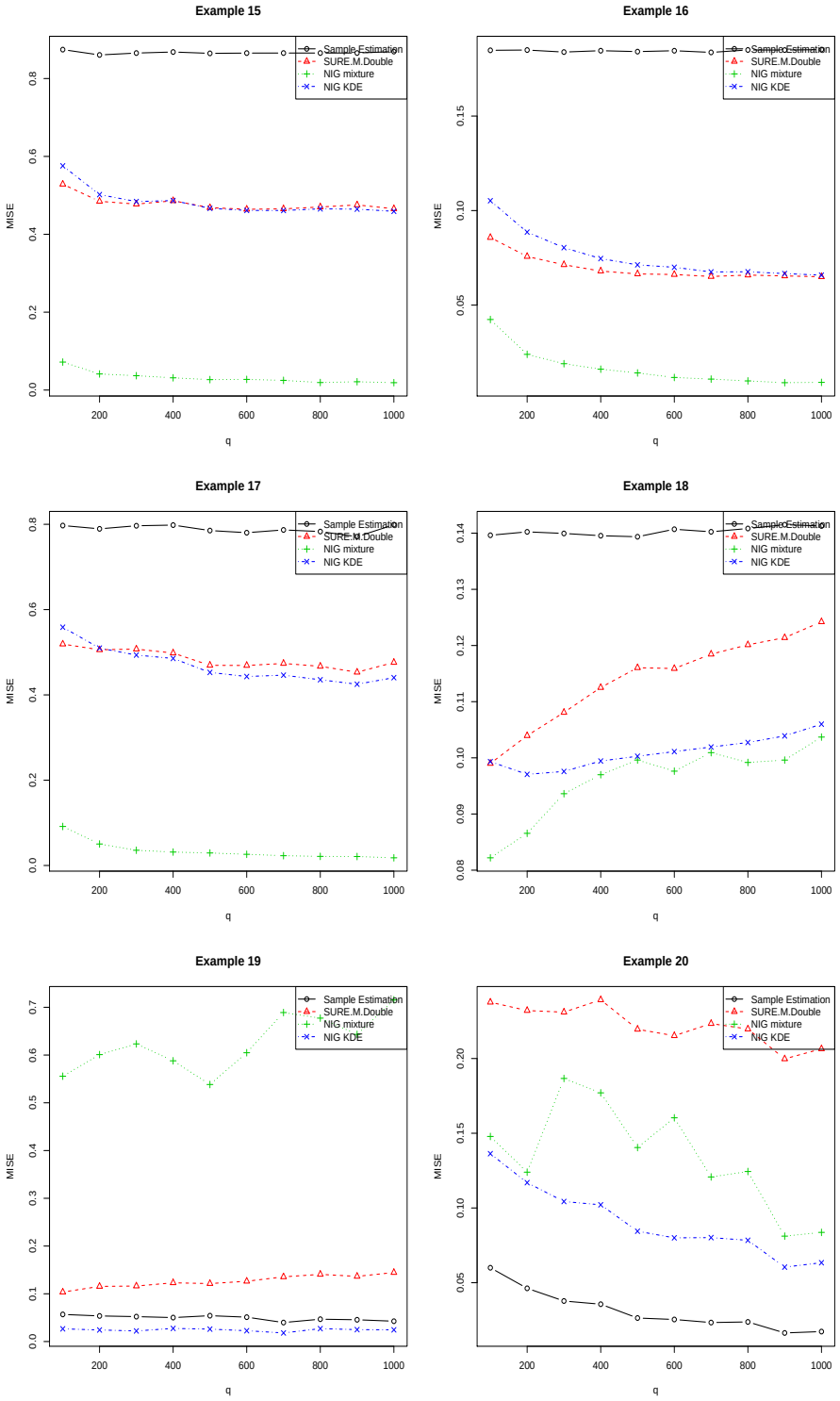


Figure 3.3: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ vs. dimension q of normal vector for Examples 15-20 of Section 3.3.1. Dimension size is $q = 100, 200, \ldots, 1000$ and number of replications is 100 for each q.



of interest, in which case

$$\mu_i = \frac{1}{50}\sum_{j=1}^{50} X_{ij} \quad \text{and} \quad \sigma_i^2 = \frac{1}{50}\sum_{j=1}^{50}(X_{ij}-\mu_i)^2, \quad i = 1, \ldots, 6033.$$

As a test of the various methods, we randomly select four subjects from the control group and use their data to estimate μ_i and σ_i^2.

To better understand the nature of the data we provide density plots in Figure 3.4. We selected columns 6, 30, 31, and 48 and used our method to estimate f_μ and f_{σ^2}. The "true" distribution of μ and σ^2 was generated from the density command in R with the default bandwidth.

To compare different methods we randomly chose all rows and 4 columns, computed the MISEs of various estimates of the three densities, and replicated this process 100 times. The MISE for each method was computed as in our simulation study. Table 3.4 shows that all methods were outperformed by $N\Gamma^{-1}$. This is not too surprising considering that Figure 3.4 shows that the densities of μ_i and σ_i^2 are well-approximated by normal and inverse gamma densities, respectively. For all algorithms, except SURE.M.Double, we replace the unknown σ_i^2 with $S_{i\cdot}^2$, which results in a loss of accuracy for those methods.

In Table 3.5 we provide results from a simulation in which we randomly selected only 1000 genes and 4 subjects. Here also the $N\Gamma^{-1}$ method performs better than the other methods.

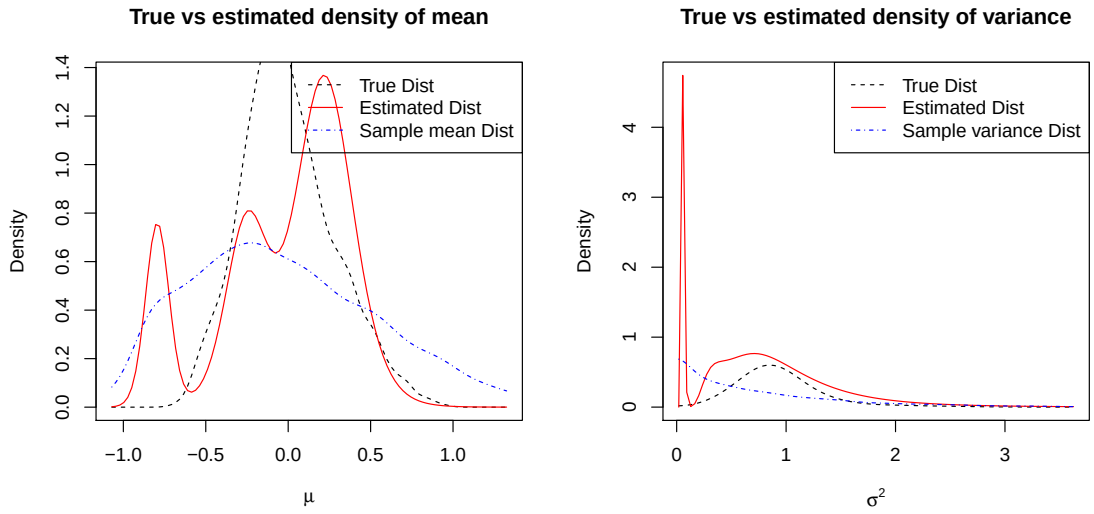
Table 3.4: Estimates of MISE for the prostate data. Each table value is an average of 100 replications. In each run 4 of 50 subjects were randomly selected and all 6033 genes were used. The total number of components used in a mixture was $k = 10$.

Different Methods	$MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$	$MISE(\hat{f}_{\mu}, f_{\mu})$	$MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$
Sample Statistics	0.6433	0.2863	0.4160
EBMLE.XKB	0.4684	0.0629	0.4160
EBMOM.XKB	0.5809	0.3640	0.4160
JS.XKB	0.6390	0.2781	0.4160
Oracle.XKB	0.7164	0.4049	0.4160
SURE.G.XKB	0.9946	0.8624	0.4160
SURE.M.XKB	1.2171	0.8844	0.4160
SURE.SG.XKB	1.9519	3.2717	0.4160
SURE.SM.XKB	1.7600	2.6570	0.4160
SURE.M.Double	4.9595	0.9289	0.7561
$N\Gamma^{-1}$ mixture	0.4040	0.1624	0.0738

Table 3.5: Estimates of MISE for the prostate data. Each table value is an average of 1000 replications. In each run 4 of 50 subjects were randomly selected and 1000 genes were randomly selected from all 6033 genes. The total number of components used in a mixture was $k = 10$.

Different Methods	$MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$	$MISE(\hat{f}_{\mu}, f_{\mu})$	$MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$
Sample Statistics	0.5841	0.2791	0.3862
EBMLE.XKB	0.4173	0.0681	0.3862
EBMOM.XKB	0.5232	0.3848	0.3862
JS.XKB	0.5785	0.2696	0.3862
Oracle.XKB	0.5847	0.3628	0.3862
SURE.G.XKB	0.8362	0.8405	0.3862
SURE.M.XKB	0.9958	0.8349	0.3862
SURE.SG.XKB	1.4758	2.4118	0.3862
SURE.SM.XKB	1.4807	2.4301	0.3862
SURE.M.Double	4.8824	1.1655	0.7390
$N\Gamma^{-1}$ mixture	0.1962	0.0286	0.0349

Figure 3.4: True vs. estimated marginal densities. The estimated marginal density of μ and σ^2 is based on $N\Gamma^{-1}$ method using only columns 6, 30, 31, and 48



4. LOCATION-SCALE DENSITY ESTIMATION USING HISTOGRAM

As mentioned in Section 1.2, the main focus of this chapter is semiparametric estimation of the joint density of μ and σ^2 in model (1.6). In this section, we discuss the method of estimating the density f_{μ,σ^2} using a bivariate histogram when f_ϵ is any known density with mean 0 and variance 1. This approach is semiparametric in nature as we assume the error density is known but f_{μ,σ^2} is unknown.

4.1 Location-Scale Density Estimation with Known Error Density

In this section, we will discuss the method of estimating f_{μ,σ^2} using a bivariate histogram for a few known choices of f_ϵ. In Section 4.1.1, we discuss a general method of estimating f_{μ,σ^2} using a bivariate histogram.

4.1.1 Modeling the Distribution of Location-Scale with a Histogram

If we assume that f_ϵ is known, with mean 0 and variance 1, then our problem boils down to estimating the joint distribution of (μ_i, σ_i^2). Choosing the best f_ϵ is not an easy task but for the moment we assume that it is known. To estimate the distribution of f_{μ,σ^2} we will use a histogram representation. Let $I_A(x)$ be defined as

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise.} \end{cases}$$

Then, the histogram approximation to the joint distribution of (μ_i, σ_i^2) can be written as

$$f_{\mu,\sigma^2}(m, v|\boldsymbol{p}_k, k) = \sum_{r=1}^{k} \frac{p_{r,k}}{A_r} I_{R_r}(m, v),$$

where $R_1, \ldots, R_k$ are k bins, and A_r is the area of R_r. It can be shown that the support of (μ, σ^2) is contained in the support of $(\bar{X}_{i\cdot}, S_{i\cdot}^2)$ where $\bar{X}_{i\cdot} = n^{-1} \sum_{j=1}^{n} X_{ij}$ is the sample mean and $S_{i\cdot}^2 = (n-1)^{-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{i\cdot})^2$ is the sample variance of the i^{th} dataset. The quantity $p_{r,k}$ is the probability that (μ_i, σ_i^2) belongs to the r^{th} bin when the total number of bins is k.

Let $\boldsymbol{X}_{i\cdot} = (X_{i1}, \ldots, X_{in})^T$ be the vector of observed replications of the true unobserved variable μ_i. Let J_{ir} be defined as

$$J_{ir} = \int \int_{R_r} \prod_{j=1}^{n} \frac{1}{\sqrt{v}} f_\epsilon \left(\frac{X_{ij} - m}{\sqrt{v}} \right) dmdv.$$

Then the likelihood for the i^{th} dataset can be written as

$$L_i(\boldsymbol{p}_k) = f(\boldsymbol{X}_{i\cdot}|\boldsymbol{p}_k, f_\epsilon, k) = \sum_{r=1}^{k} \frac{p_{r,k}}{A_r} J_{ir}.$$

Let $\mathbf{X}$ denotes the all $q \times n$ observations, $\boldsymbol{X}_{1\cdot}, \ldots, \boldsymbol{X}_{q\cdot}$. Then the complete data log-likelihood for all q datasets can be written as

$$l(\boldsymbol{p}_k) = \log(f(\mathbf{X}|\boldsymbol{p}_k, f_\epsilon, k)) = \sum_{i=1}^{q} \log\left(L_i(\boldsymbol{p}_k)\right).$$

Our objective is to maximize the posterior distribution

$$h(\boldsymbol{p}_k) \stackrel{\text{def}}{=} f(\boldsymbol{p}_k|\mathbf{X}, f_\epsilon, k) \propto \prod_{i=1}^{q} L_i(\boldsymbol{p}_k)\pi(\boldsymbol{p}_k|\boldsymbol{\alpha}), \tag{4.1}$$

where $\pi(\boldsymbol{p}_k|\boldsymbol{\alpha})$ is the prior for $\boldsymbol{p}_k$, which we take to be a non-informative Dirichlet prior with $\boldsymbol{\alpha} = (1/2, \ldots, 1/2)$.

Since the Dirichlet is not a conjugate prior, we need to use the Metropolis-Hastings algorithm to approximate the posterior distribution of $\boldsymbol{p}_k$.

4.1.2 Different Choices for the Distribution of Scaled Error

The method described in Section 4.1.1 is similar for any choice of a scaled error distribution. The only thing that varies is the form of J_{ir}. In this section, we compute J_{ir} for a few popular distributions.

4.1.2.1 Standard Normal Scaled Error Distribution

As mentioned before, in most of the early literature on measurement error it was assumed that the error density was normal. Carroll and Hall [1988], Fan [1991], and Fan [1992] assumed normality of error. In the more recent work of Staudenmayer et al. [2008], normality of the error density was also assumed. Sarkar et al. [2014] used the normal distribution as one of the possible choices for the scaled error density in the LSRE model.

If f_ϵ is standard normal and R_r is the rectangle $[a_{s-1}, a_s] \times [b_{t-1}, b_t]$, then

$$\begin{aligned} J_{ir} &= (2\pi)^{-n/2} \int_{b_{t-1}}^{b_t} \int_{a_{s-1}}^{a_s} v^{-n/2} e^{-(n-1)S_{i\cdot}^2/2v - n(\bar{X}_{i\cdot}-m)^2/2v} dm dv \\ &= \frac{1}{\sqrt{n(2\pi)^{n-1}}} \int_{b_{t-1}}^{b_t} e^{-(n-1)S_{i\cdot}^2/2v} \frac{1}{v^{(n-1)/2}} \left[\Phi\left(\frac{a_s - \bar{X}_{i\cdot}}{\sqrt{v/n}} \right) - \Phi\left(\frac{a_{s-1} - \bar{X}_{i\cdot}}{\sqrt{v/n}} \right) \right] dv. \end{aligned}$$

This integration does not have a closed form solution and we need numerical integration to calculate J_{ir}.

4.1.2.2 Uniform Scaled Error Distribution

The uniform distribution could be used in the case where we know that the scaled error density is short-tailed. As discussed in Hart and Cañette(2012), using a long-tailed density in such cases could potentially lead to improper estimation of the density of interest, f_{μ,σ^2}.

Let f_ϵ be a uniform density between $-c$ and c. If $c = \sqrt{3}$, then the variance of f_ϵ is 1. In this case

$$\begin{aligned} J_{ir} &= \int_{b_{t-1}}^{b_t} \int_{a_{s-1}}^{a_s} (2c\sqrt{v})^{-n} I_{(X_{i(n)}-\sqrt{v}c, X_{i(1)}+\sqrt{v}c)}(m) dm dv \\ &= \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{s-1}}^{a_s} (2cs)^{-n} I_{(X_{i(n)}-sc, X_{i(1)}+sc)}(m)(2s) dm ds \quad (\text{where} \quad s = \sqrt{v}) \\ &= 2^{-n+1} c^{-n} \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{s-1}}^{a_s} s^{-n+1} I_{(X_{i(n)}-sc, X_{i(1)}+sc)}(m) dm ds. \end{aligned}$$

Let the maximum and minimum of the i^{th} dataset be $X_{i(n)}$ and $X_{i(1)}$, respectively. Geometrically, we can visualize this as a 2-D plane of (m, s). J_{ir} is 0 for any rectangle if it lies entirely below either of the lines $s = (m - X_{i(1)})/c$ or $s = (X_{i(n)} - m)/c$.

How two lines intersect with a rectangle creates a number of subcases. In this case, though, we do not need numerical integration. Instead, we need to consider the subcases separately. In Appendix A we considered all possible cases to compute the integral J_{ir}.

4.1.2.3 Other Possible Scaled Error Distributions

We can likewise compute J_{ir} for any known f_ϵ that has mean 0 and standard deviation 1. If we need a long-tailed error density with tails heavier than the normal then we could choose Student's t-distribution with appropriate degrees of freedom. The t-distribution simply needs to be scaled to have variance 1 and the degrees of freedom have to be greater than 2 to ensure finite variance. For this reason, we can not use the Cauchy distribution as f_ϵ directly, unless we change the basic assumption about f_ϵ to have median 0 and IQR 1. A mixture of normal densities is also a possible candidate in case we need a multimodal scaled error distribution.

Though it is possible to estimate error density parameters such as the degrees of freedom in the case of the t-distribution or the mixing probabilities in the case of a mixture of normal densities, we try to avoid this since then we need to compute J_{ir} in each iteration of MCMC, and this quantity involves $q \times k$ numerical integrations. This significantly slows down the estimation algorithm. So, instead of trying to estimate error density parameters, we will try to estimate the error density nonparametrically, as discussed in Section 4.2.

4.1.3 Algorithm to Estimate the Bin Probabilities

We wish to find the maximum likelihood estimate of $\boldsymbol{p}_k$ in order to have a starting value for our MCMC procedure. Once we calculate J_{ir} for each rectangular bin, we can write down the likelihood. Now, to find the MLE of $\boldsymbol{p}_k$ we need to maximize $l(\boldsymbol{p}_k)$ subject to the constraints $\boldsymbol{p}_k \geq \boldsymbol{0}$ and $\sum_{r=1}^{k} p_{r,k} = 1$. By reparameterizing as follows we can automatically ensure the constraints. $p_{r,k} = e^{a_r} / \sum_{s=1}^{k} e^{a_s}$, $r = 1, \ldots, k$.

To ensure identifiability, we set $a_1 = 0$. Now, we can rewrite the likelihood as a function of $a_2, \ldots, a_k$, which is easy to maximize using the Newton-Raphson method. The log-likelihood as a function of $\boldsymbol{a} = (a_1, \ldots, a_k)$ is

$$l(\boldsymbol{a}) = \sum_{i=1}^{q} \left\{ \log \left(\sum_{r=1}^{k} e^{a_r} {A_r}^{-1} J_{ir} \right) - \log \left(\sum_{r=1}^{k} e^{a_s} \right) \right\}.$$

Derivatives of the log-likelihoods are

$$\frac{\partial l(\boldsymbol{a})}{\partial a_m} = \sum_{i=1}^{q} \left\{ \frac{{A_m}^{-1} J_{im} e^{a_m}}{\sum_{r=1}^{k} {A_r}^{-1} J_{ir} e^{a_r}} \right\} - q \frac{e^{a_m}}{\sum_{r=1}^{k} e^{a_r}},$$

$$\frac{\partial^2 l(\boldsymbol{a})}{\partial a_m^2} = \sum_{i=1}^{q} \left\{ \frac{\sum_{r \in \{1,\ldots,k\} \setminus \{m\}} {A_r}^{-1} {A_m}^{-1} J_{ir} J_{im} e^{a_r} e^{a_m}}{\left(\sum_{r=1}^{k} {A_r}^{-1} J_{ir} e^{a_r} \right)^2} \right\} - q \frac{\sum_{r \in \{1,\ldots,k\} \setminus \{m\}} e^{a_r} e^{a_m}}{\left(\sum_{r=1}^{k} e^{a_r} \right)^2},$$

$$\text{for} \quad m = 2, \ldots, k,$$

$$\frac{\partial^2 l(\boldsymbol{a})}{\partial a_m \partial a_s} = \sum_{i=1}^{q} \left\{ - \frac{{A_m}^{-1} {A_s}^{-1} J_{im} J_{is} e^{a_m} e^{a_s}}{\left(\sum_{r=1}^{k} {A_r}^{-1} J_{ir} e^{a_r} \right)^2} \right\} + q \frac{e^{a_m} e^{a_s}}{\left(\sum_{r=1}^{k} e^{a_r} \right)^2}$$

$$\text{for} \quad m, s = 2, \ldots, k, \text{ and } m \neq s.$$

After starting with a reasonable initial value $\boldsymbol{a}^{(0)}$, the Newton-Raphson iterates converge to a value that we call $\boldsymbol{a}^{MLE}$. Define $\boldsymbol{a}^{(0)} = \log(0.9 {\boldsymbol{p}_k}^{naive} + 0.1(1/k, \ldots, 1/k))$, where, ${\boldsymbol{p}_k}^{naive}$ are the bivariate histogram probabilities of $(\bar{X}_i, S_i^2),\ i = 1, \ldots, q$. The iterates of the Newton-Raphson scheme are

$$\boldsymbol{a}^{(t)} = \boldsymbol{a}^{(t-1)} - \left\{ \frac{\partial^2 l(\boldsymbol{a})}{\partial \boldsymbol{a}^2} \right\}^{-1}_{\boldsymbol{a} = \boldsymbol{a}^{(t-1)}} \left\{ \frac{\partial l(\boldsymbol{a})}{\partial \boldsymbol{a}} \right\}_{\boldsymbol{a} = \boldsymbol{a}^{(t-1)}}, \quad t = 1, 2, \ldots.$$

After maximizing the log-likelihood with respect to the a_r's for $r = 2, \ldots, k$, we find $\boldsymbol{a}^{MLE}$ or ${\boldsymbol{p}_k}^{MLE}$, which we will use as a starting value in performing MCMC. The posterior density $h(\boldsymbol{p}_k)$ is defined by (4.1). Let k and s denote the length of the probability vector

and the number of MCMC iterations respectively. Let f be a proposal density with support $(0, \infty)$. The conditional density q_r is defined by

$$q_r(\tilde{\boldsymbol{p}}_k|\boldsymbol{p}_k) = \begin{cases} f\left(\tilde{p}_{r,k}\frac{1-p_{r,k}}{1-\tilde{p}_{r,k}}\right)\frac{(1-p_{r,k})}{(1-\tilde{p}_{r,k})^2}, & \tilde{p}_{t,k} = \frac{p_{t,k}(1-\tilde{p}_{r,k})}{(1-p_{r,k})} \text{ for } t \in \{1,\ldots,k\}\setminus\{r\} \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 1 MCMC algorithm

1: Initialize $l = 0$ and start the MCMC chain with $\boldsymbol{p}^{(0)} = \boldsymbol{p}_k{}^{MLE}$.
2: Initialize $r = 1$.
3: Generate Q from proposal density f with mean $p_{r,k}$. Then define $\tilde{p}_{r,k} = \frac{Q}{(1-p_{r,k}+Q)}$ and $\tilde{p}_{t,k} = \frac{p_{t,k}}{(1-p_{r,k}+Q)}$ for $t \in \{1,\ldots,k\}\setminus\{r\}$.
4: Set $\boldsymbol{p}_k{}^{(l+1)} = \tilde{\boldsymbol{p}}_k$ with probability $\min\left(1, \frac{h(\tilde{\boldsymbol{p}}_k)q_r(\boldsymbol{p}_k|\tilde{\boldsymbol{p}}_k)}{h(\boldsymbol{p}_k)q_r(\tilde{\boldsymbol{p}}_k|\boldsymbol{p}_k)}\right)$, and otherwise set $\boldsymbol{p}_k{}^{(l+1)} = \boldsymbol{p}_k{}^{(l)}$.
5: $r = r + 1$ and if $r < k$ goto step 3.
6: $l = l + 1$ and if $l < s$ goto step 2.

The proportion $\frac{q_r(\tilde{\boldsymbol{p}}_k|\boldsymbol{p}_k)}{q_r(\boldsymbol{p}_k|\tilde{\boldsymbol{p}}_k)}$ is always positive if we start from a feasible $\boldsymbol{p}_k$ that satisfies all constraints. Here, we are not trying to update $\boldsymbol{p}_k$ all at once. Since the dimension of $\boldsymbol{p}_k$ is large, updating with the Metropolis-Hastings algorithm all at once is difficult. So, we are updating $\boldsymbol{p}_k$ componentwise. Finally, we calculate $\hat{\boldsymbol{p}}_k$ by taking the average of all iterations, where $\hat{\boldsymbol{p}}_k$ denotes estimated bin probabilities with the number of bins equal to k.

4.1.4 Model Selection

There are a few issues we need to address before fitting the model:

(i) Selecting the support of the histogram.

(ii) Selecting the appropriate error density.

(iii) Selecting the number of bins.

4.1.4.1 Selecting the Support of the Histogram

As we already discussed, the support of (μ, σ^2) is contained in the support of $(\bar{X}_{i\cdot}, S_{i\cdot}^2)$, but in reality, using the latter support is quite conservative. This means that most of the bin probabilities will be 0. To overcome this problem, we will use a two-step procedure. In the first step we run the algorithm on $\left[\min(\overline{\boldsymbol{X}}), \max(\overline{\boldsymbol{X}})\right] \times [\min(\boldsymbol{S}^2), \max(\boldsymbol{S}^2)]$. The min and max is defined in Section 3.3.1. Then we get rid of those bins where the estimated probability is less than some small positive quantity. During the second step we further divide the bins that have bigger probabilities. There are many ways to define bins that are rectangular boxes, and it is not clear how to do this in an optimal fashion. A simple solution is to divide the range of $\bar{X}_{i\cdot}$ into k_1 equal parts and the range of $S_{i\cdot}^2$ into k_2 equal parts. So, $k = k_1 \times k_2$ and the sides of any rectangle are parallel to the axes. For simplicity we choose $k_1 = k_2$.

4.1.4.2 Selecting an Error Density and Number of Bins

As we discussed before, selecting the correct error density is not easy. In this section, we propose using a Bayes factor to select the error density and the optimum number of bins for a given error density. Let $M_{(k,f_\epsilon)}$ denote the model under consideration for a given combination of k and f_ϵ. The quantity $\boldsymbol{p}_k$ is the vector of probabilities of length k. As a prior on k given that the error model is f_ϵ we use $\pi(k|f_\epsilon) \propto 1/k$. The prior on the model $M_{(k,f_\epsilon)}$ is $\pi(M_{(k,f_\epsilon)}) = \pi(k|f_\epsilon)\pi(f_\epsilon)$, and we assume that $\pi(f_\epsilon)$ is the same for all f_ϵ. Let $m_{(k,f_\epsilon)}$ denote the marginal of the model $M_{(k,f_\epsilon)}$ which is defined as

$$m_{(k,f_\epsilon)} = \int_{[0,1]^k \cap \sum_{r=1}^k p_{r,k}=1} f(\mathbf{X}|\boldsymbol{p}_k, k, f_\epsilon)\pi(\boldsymbol{p}_k|k, f_\epsilon)d\boldsymbol{p}_k$$

and

$$\pi(\boldsymbol{p}_k, k, f_\epsilon|\mathbf{X}) \propto f(\mathbf{X}|\boldsymbol{p}_k, k, f_\epsilon)\pi(\boldsymbol{p}_k|k, f_\epsilon)\pi(k|f_\epsilon)\pi(f_\epsilon).$$

Let $\pi(M_{(k,f_\epsilon)}|\mathbf{X})$ be the posterior probability of model $M_{(k,f_\epsilon)}$ is defined as

$$\pi(M_{(k,f_\epsilon)}|\mathbf{X}) \propto m_{(k,f_\epsilon)}\pi(M_{(k,f_\epsilon)}).$$

The posterior probability of normal error density is

$$\pi(f_\epsilon = N(0,1)|\mathbf{X}) = \sum_{k\in\mathbb{N}^2} \pi(M_{(k,f_\epsilon=N(0,1))}|\mathbf{X}).$$

To compare two candidates for the error model, say normal vs. uniform, we can compute the following posterior odds ratio as

$$\frac{\pi(f_\epsilon = N(0,1)|\mathbf{X})}{\pi(f_\epsilon = U(-\sqrt{3},\sqrt{3})|\mathbf{X})}.$$

Now, we can select the error density, denoted by $\tilde{f}_\epsilon$, that maximizes $\pi(f_\epsilon|\mathbf{X})$. The posterior distribution of k given the selected error density is

$$\pi(k|\mathbf{X},\tilde{f}_\epsilon) = \frac{m_{(k,\tilde{f}_\epsilon)}\pi(k|\tilde{f}_\epsilon)}{\sum_{r\in\mathbb{N}^2} m_{(r,\tilde{f}_\epsilon)}\pi(r|\tilde{f}_\epsilon)}.$$

Let $\hat{k}$ denote the optimum k that maximizes the posterior distribution of k given $\tilde{f}_\epsilon$. Then

$$\hat{k} = \arg\max_k \left(\log m_{(k,\tilde{f}_\epsilon)} - \log k\right).$$

Also, we can find an optimum k using criteria such as AIC or BIC. These are defined by

$$\hat{k}_{AIC} = \arg\max_k \left(l(\hat{\boldsymbol{p}}_k) - k\right)$$
$$\hat{k}_{BIC} = \arg\max_k \left(l(\hat{\boldsymbol{p}}_k) - \frac{k}{2}\log(q)\right).$$

To calculate $m_{(k,f_\epsilon)}$ we can use Laplace approximation or importance sampling. The Laplace approximation is

$$\log(m_{(k,f_\epsilon)}) \approx \log(f(\mathbf{X}|\breve{\boldsymbol{p}}_k, k, f_\epsilon)) + \log(\pi(\breve{\boldsymbol{p}}_k|k)) + (k/2)\log(2\pi) - (1/2)\log|\Sigma| \\ -(k/2)\log(q),$$

where,

$$\breve{\boldsymbol{p}}_k = \arg\max_{\boldsymbol{p}_k} \log(h(\boldsymbol{p_k})) \quad \text{and} \quad \Sigma = \left\{ \frac{\partial^2 \left\{ -\frac{1}{q}\log(h(\boldsymbol{p_k})) \right\}}{\partial {\boldsymbol{p}_k}^2} \right\}_{\boldsymbol{p}_k = \breve{\boldsymbol{p}}_k}.$$

Another way to estimate $m_{(k,f_\epsilon)}$ is importance sampling. We have

$$\begin{aligned} m_{(k,f_\epsilon)} &= \int_{[0,1]^k \cap \sum_{r=1}^k p_{r_k}=1} f(\mathbf{X}|\boldsymbol{p}_k, k, f_\epsilon)\pi(\boldsymbol{p}_k|k)d\boldsymbol{p}_k \\ &= \int_{[0,1]^k \cap \sum_{r=1}^k p_{r,k}=1} f(\mathbf{X}|\boldsymbol{p}_k, k, f_\epsilon)\frac{\pi(\boldsymbol{p}_k|k)}{g(\boldsymbol{p}_k|k)}g(\boldsymbol{p}_k|k)d\boldsymbol{p}_k \\ &\approx \frac{1}{N}\sum_{l=1}^{N} f(\mathbf{X}|\boldsymbol{p}_k^{(l)}, k, f_\epsilon)\frac{\pi(\boldsymbol{p}_k^{(l)}|k)}{g(\boldsymbol{p}_k^{(l)}|k)}, \end{aligned}$$

where, ${\boldsymbol{p}_k}^{(1)}, \ldots {\boldsymbol{p}_k}^{(N)}$ are random draws from $g(\boldsymbol{p}_k|k)$.

The choice of $g(\boldsymbol{p}_k|k)$ is tricky. We choose a $k-1$ dimensional multivariate normal with posterior mean and posterior variance calculated from MCMC iterations. We randomly choose a number $t \in \{1, \ldots, k\}$ at each simulation and then for $r \in \{1, \ldots, k\} \setminus \{t\}$, $p_{r,k}$ follows a univariate normal with mean $\hat{p}_{r,k}$ and standard deviation calculated from all MCMC iterations. Then define $p_{t_k} = 1 - \sum_{r \in \{1,\ldots,k\} \setminus \{t\}} p_{r,k}$. We will accept the simulated $\boldsymbol{p}_k$ if $\boldsymbol{p}_k \geq \mathbf{0}$. After generating N simulations of $\boldsymbol{p}_k$ we can estimate $m_{(k,f_\epsilon)}$. The distribution g for randomly

selected t is

$$g_t(\boldsymbol{p}_k|k) = \begin{cases} \prod_{r\in\{1,\ldots,k\}\setminus\{t\}} \frac{1}{sd(p_{r,k})}\varphi\left(\frac{p_{r,k}-\hat{p}_{r,k}}{sd(p_{r,k})}\right), & p_{t,k} = 1 - \sum_{r\in\{1,\ldots,k\}\setminus\{t\}} p_{r,k}, \\ & p_{1,k},\ldots,p_{k,k} \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where φ is the standard normal density, and $\hat{p}_{r,k}$ and $sd(p_{r,k})$ denote the mean and standard deviation, respectively of the posterior density of $p_{r,k}$ calculated from MCMC iterations.

Here, instead of using one single proposal density g for importance sampling we use a combination of proposal densities $g_1,\ldots,g_k$ and in each iteration we are choosing one of them randomly.

4.2 Nonparametric Method for Estimating the Error Density in LSRE Model

To perhaps better infer the error density we can take a non-parametric approach. Consider data $X_{ij}, i = 1,\ldots,q, j = 1,\ldots,n$, and suppose that

$$X_{ij} = \mu_i + \sigma_i\epsilon_{ij}, \quad i = 1,\ldots,q, \quad j = 1,\ldots,n.$$

We can rewrite this equation if $n \geq 4$. For $j \neq l$

$$\log(|X_{ij} - X_{il}|) = \log(\sigma_i) + \log(|\epsilon_{ij} - \epsilon_{il}|).$$

Suppose, $E(\log|\epsilon_1 - \epsilon_2|) = c$. Then for $j \neq l$ we have

$$\log(|X_{ij} - X_{il}|) = \log(\sigma_i) + c + (\log(|\epsilon_{ij} - \epsilon_{il}|) - c).$$

This is an LRE model, and Hart and Cañette [2011] describe a way to consistently estimate the density of ϵ_{ij} under very general conditions. Let us call this minimum distance estimator $\hat{f}_\epsilon$. We can substitute $\hat{f}_\epsilon$ for f_ϵ in the methodology described in Section 4.2 and therefore

find an estimate of f_{μ,σ^2}.

4.3 Simulation Study

The mean integrated squared error (MISE) measures the overall accuracy of estimating f_{μ,σ^2} with a bivariate histogram and we can estimate the MISE based on a Monte Carlo method using B simulated datasets. MISE and its estimate are defined as

$$\begin{aligned} MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2}) &= \int_{\mathbb{R}^+}\int_{\mathbb{R}} E\Big(f_{\mu,\sigma^2}(m,v) - \hat{f}_{\mu,\sigma^2}(m,v)\Big)^2 dmdv \\ &\approx B^{-1}\sum_{b=1}^{B}\sum_{r=1}^{k}\frac{1}{A_r}\left(\int_{R_r}\int f_{\mu,\sigma^2}(m,v)dmdv - \int_{R_r}\int \hat{f}^b_{\mu,\sigma^2}(m,v)dmdv\right)^2, \end{aligned}$$

where $R_r,\ r = 1,\ldots,k$, denotes the bivariate class-intervals which was used to estimate f_{μ,σ^2} and $\hat{f}^b_{\mu,\sigma^2}$ denotes the estimated density of f_{μ,σ^2} based on the b^{th} simulated dataset. We will use MISE to measure performance of this method for a wide variety of choices for f_{μ,σ^2}.

To our knowledge there exist no competitors in the literature for our method of estimating the density of f_{μ,σ^2}. A possible means of estimating f_{μ,σ^2} would be to estimate (μ_i, σ_i^2), $i = 1,\ldots,q$, and to then compute a histogram estimator from the estimates as if they were the true values of (μ_i, σ_i^2). We will call such estimates "plug-in histogram estimates." For these "plug-in histogram estimates," $\int_{R_r}\int \hat{f}_{\mu,\sigma^2}(m,v)dmdv$ is simply the proportion of estimated mean and variance pairs that lie in the rectangle R_r. There are a few articles, including Xie et al. [2012] and Weinstein et al. [2018], which address the problem of estimating μ_i when σ_i^2 is known. Jing et al. [2016] further extended the work of Xie et al. [2012] and estimated both μ_i and σ_i^2. The estimators of Xie et al. [2012] defined by their expressions (7.1), (7.2), (7.3), (4.2), and (5.1) will be called EBMLE.XKB, EBMOM.XKB, JS.XKB, SURE.G.XKB, and SURE.M.XKB, respectively. The methods which were referred to as SURE.M.Double can be found in expressions (11-12) of Jing et al. [2016]. Weinstein et al. [2018] developed the group-linear algorithm, which is referred to here as GL.WMBZ. All the above estimates assume that f_ϵ is standard normal.

We simulated data from model (1.6) for different choices of f_{μ,σ^2}. For each (μ_i, σ_i^2) pair,

there are $n = 4$ replications. Only the observations X_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, n$, are used to find the estimates of f_{μ,σ^2}. The ensuing tables compare the approximated MISEs of various methods. We also consider "plug-in histogram estimates" based on the methods of Xie et al. [2012], Jing et al. [2016], and Weinstein et al. [2018]. Since Xie et al. [2012] and Weinstein et al. [2018] assumed that $\sigma_1^2, \ldots, \sigma_q^2$ are known, we use $S_{1\cdot}^2, \ldots, S_{q\cdot}^2$ in our "plug-in histogram estimates" for these methods. We may construct a "plug-in histogram estimate" based on the true (μ_i, σ_i^2) pairs, which we call the true empirical histogram estimate.

4.3.1 Performance of Histogram Estimate for Normal and Uniform Error

In this section, data are generated from model (1.6). For each example it is assumed that f_ϵ is either standard normal, $N(0,1)$ (Case 1) or a uniform density between $-\sqrt{3}$ and $\sqrt{3}$, $U(-\sqrt{3}, \sqrt{3})$ (Case 2).

Example 21. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,3)$ and $\sigma^2 \sim IG(5,2)$. When $f_\epsilon \sim N(0,1)$, Figure 4.1 shows SURE.M.Double performs better than the other methods as it assumes a normal density for μ and Inverse-gamma density for σ^2, which is the true model. The bivariate histogram method does not perform better than the other methods in both cases as all other "plug-in histogram estimates" assume that μ is normal. In Figure 4.4, SURE.M.Double does not perform well when f_ϵ is uniform.*

Example 22. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim N(0,3)$ and $\sigma^2 \sim G(9,3)$. This example is quite similar to Example 21. Except for SURE.M.Double, the other methods perform similarly in estimating the joint density f_{μ,σ^2} in both cases. SURE.M.Double outperforms other methods when f_ϵ is standard normal but performs poorly compared to other methods when f_ϵ is standard uniform. The bivariate histogram also does not performs better than other "plug-in histogram estimates" as the true μ distribution is normal.*

Example 23. *The density f_{μ,σ^2} is such that μ and σ^2 are independent and follow a bivariate histogram with class boundaries $(-2, -1.2, -0.4, 0.4, 1.2, 2)^T$ and $(0.01, 0.8, 1.6, 2.4, 3.2, 4)^T$*

for μ and σ^2, respectively. The probability vector $\boldsymbol{p}_k$ is drawn from a Dirichlet distribution with parameter $25^{-1}\mathbf{1}_{25}$, where $\mathbf{1}_{25}$ denotes a vector of 1s of length 25. In this case, the true underlying model is a histogram, which is not a smooth distribution, and hence our histogram estimates perform better than all other methods in both cases for estimating f_{μ,σ^2}, f_μ, and f_{σ^2}.

Example 24. *The density f_{μ,σ^2} is such that μ and σ^2 are independent with $\mu \sim 0.5U(1,2) + 0.5U(4,5)$ and $\sigma^2/n \sim U(0.1,1)$. In this example, the distribution of f_μ is bimodal. The histogram estimate outperforms other methods in both cases as other "plug-in estimates" discussed in Xie et al. [2012] and Jing et al. [2016] work well only for unimodal densities and the group-linear algorithms discussed in Weinstein et al. [2018] do not perform well when μ and σ^2 are independent.*

Example 25. *The density f_{μ,σ^2} is such that $\mu \sim Inv-\chi_4^2$ and $log(\sigma^2)|\mu \sim N(log(\mu), 0.5^2)$. $Inv-{\chi_4}^2$ denotes an inverse-chi square distribution with four degrees of freedom. As μ and σ^2 are not independent, SURE based methods do not perform well compared to histogram and group-linear methods.*

Example 26. *The density f_{μ,σ^2} is such that $\mu \sim \Gamma(1, 0.5)$ and $\sigma^2|\mu \sim \chi_\mu^2 + 0.1$, where ${\chi_\mu}^2$ denotes a chi-square density with μ degrees of freedom. Like Example 25, μ and σ^2 are not independent, so the histogram and group-linear methods perform better than other SURE methods.*

Figure 4.1: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is standard normal. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

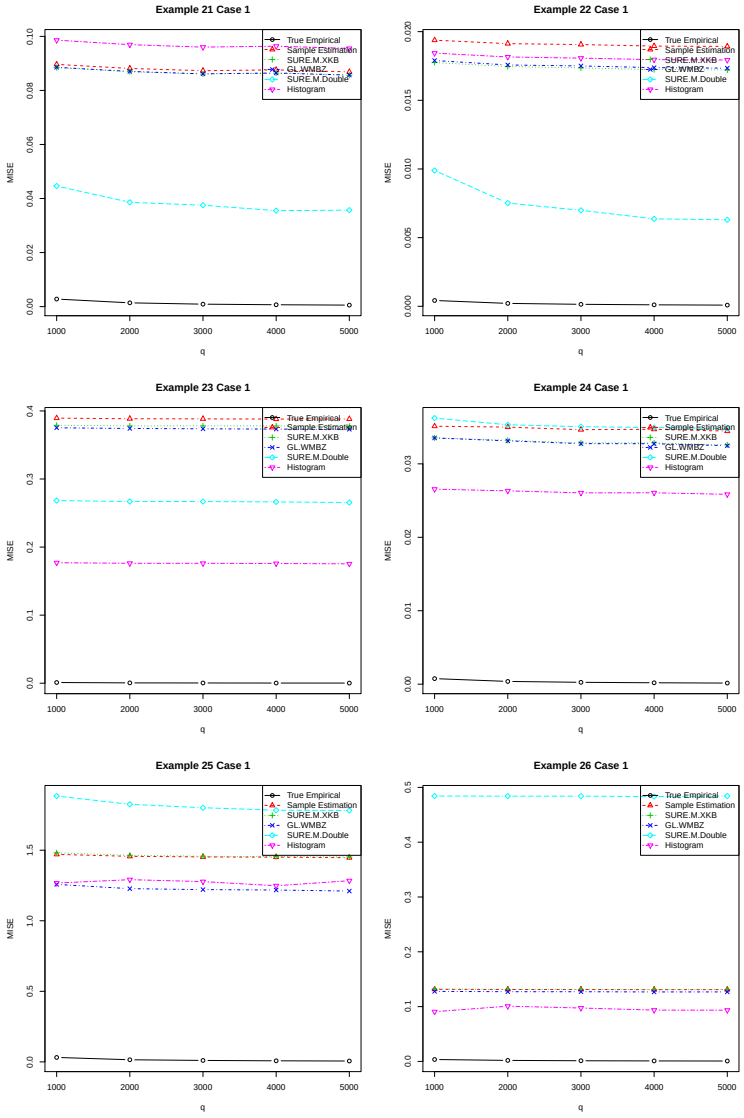


Table 4.1: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim N(0,1)$, and f_{μ,σ^2} defined by Examples 21-26 of Section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different	MISE of f_{μ,σ^2}					
Methods	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0013	0.0002	0.0005	0.0003	0.0140	0.0016
Sample Statistics	0.0879	0.0191	0.3885	0.0348	1.4554	0.1312
EBMLE.XKB	0.0868	0.0176	0.3741	0.0327	1.2453	0.1300
EBMOM.XKB	0.0868	0.0176	0.3740	0.0327	1.4032	0.1303
JS.XKB	0.0881	0.0191	0.3832	0.0348	1.3821	0.1273
SURE.G.XKB	0.0867	0.0174	0.3757	0.0330	1.4537	0.1310
SURE.M.XKB	0.0867	0.0174	0.3781	0.0330	1.4619	0.1315
GL.WMBZ	0.0868	0.0175	0.3739	0.0329	1.2268	0.1271
SURE.M.Double	0.0384	0.0074	0.2668	0.0353	1.8150	0.4839
Histogram	0.0967	0.0181	0.1761	0.0262	1.2736	0.0952

Figure 4.2: Estimated $MISE(\hat{f}_\mu, f_\mu)$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is standard normal. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

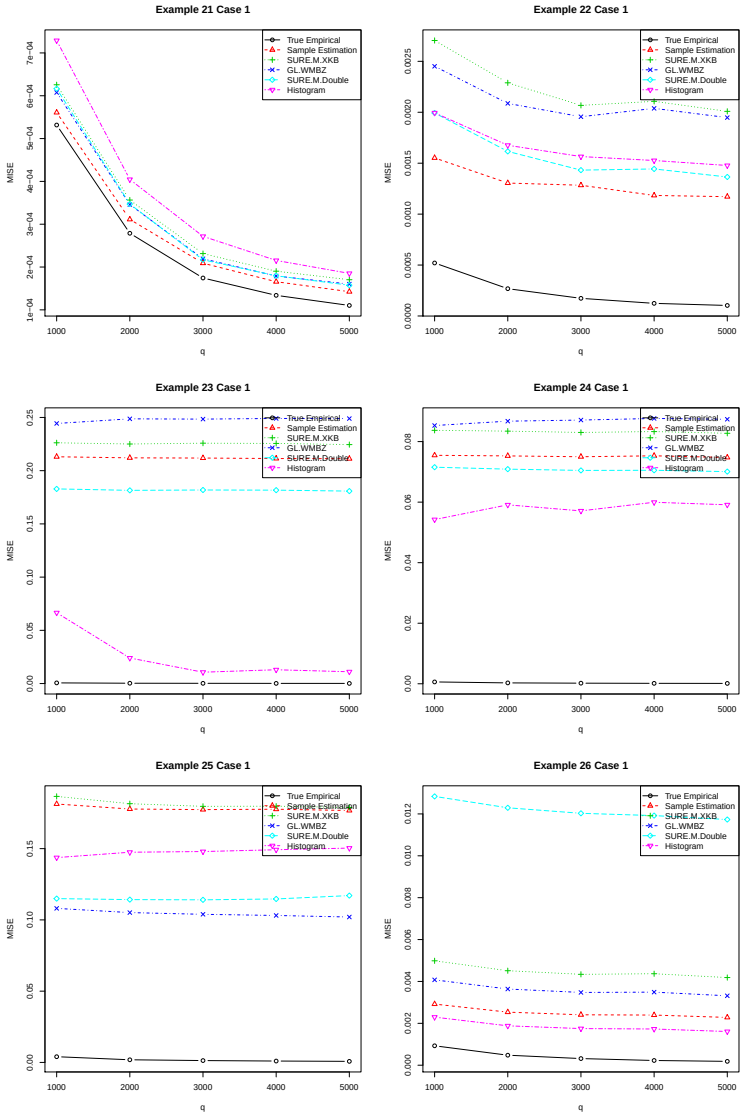


Table 4.2: Estimated $MISE(\hat{f}_\mu, f_\mu)$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim N(0, 1)$, and f_{μ,σ^2} defined by Examples 21-26 of Section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different	MISE of f_μ					
Methods	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0002	0.0002	0.0003	0.0003	0.0018	0.0004
Sample Statistics	0.0003	0.0013	0.2119	0.0752	0.1782	0.0025
EBMLE.XKB	0.0003	0.0011	0.1927	0.0763	0.1127	0.0037
EBMOM.XKB	0.0003	0.0013	0.1968	0.0772	0.1483	0.0032
JS.XKB	0.0003	0.0004	0.1883	0.0697	0.1046	0.0019
SURE.G.XKB	0.0003	0.0022	0.2129	0.0832	0.1767	0.0025
SURE.M.XKB	0.0003	0.0022	0.2254	0.0833	0.1813	0.0045
GL.WMBZ	0.0003	0.0021	0.2479	0.0868	0.1044	0.0036
SURE.M.Double	0.0003	0.0016	0.1817	0.0707	0.1150	0.0122
Histogram	0.0004	0.0016	0.0251	0.0579	0.1478	0.0019

Figure 4.3: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is standard normal. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

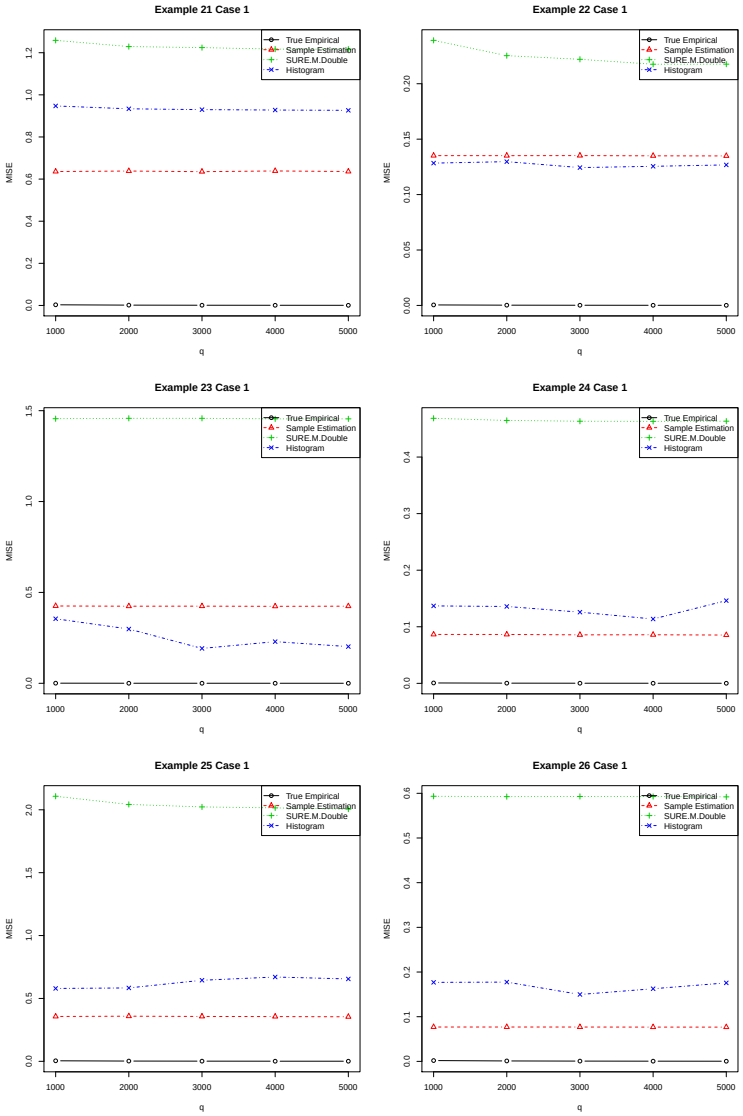


Table 4.3: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim N(0, 1)$, and f_{μ,σ^2} defined by Examples 21-26 of Section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different	MISE of f_{σ^2}					
Methods	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0014	0.0002	0.0003	0.0003	0.0019	0.0009
Sample Statistics	0.6368	0.1350	0.4244	0.0860	0.3564	0.0768
SURE.M.Double	1.2293	0.2243	1.4563	0.4646	2.0387	0.5927
Histogram	0.9330	0.1269	0.2553	0.1317	0.1317	0.1685

Figure 4.4: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is uniform. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

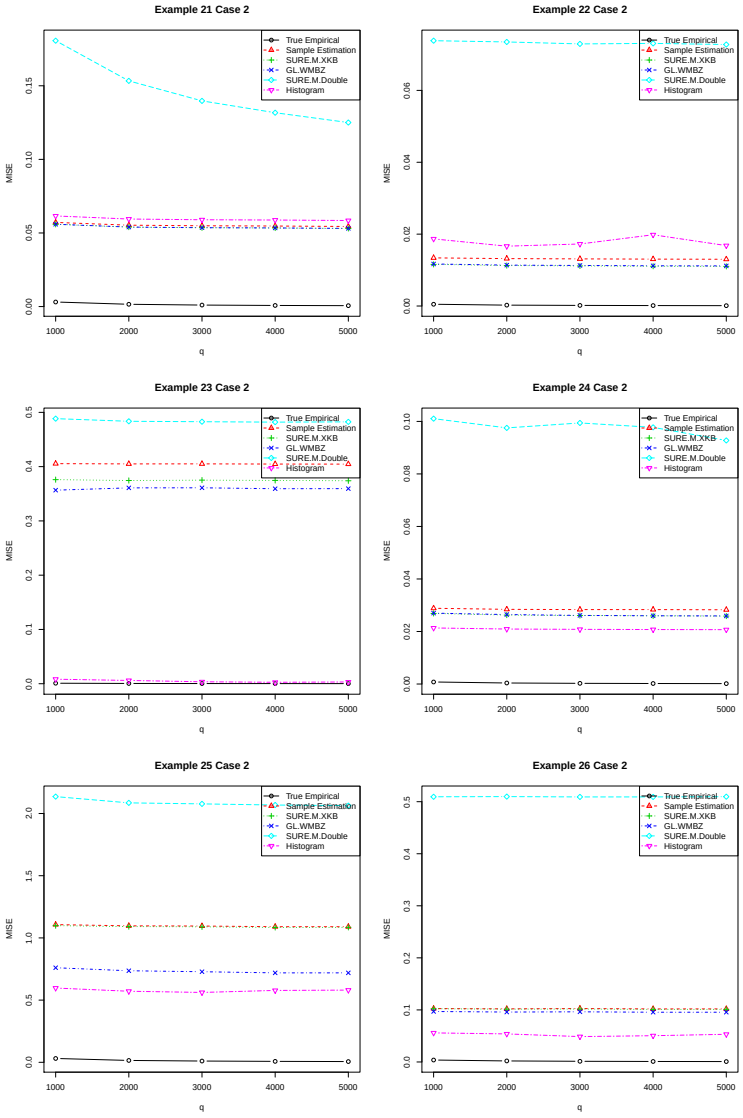


Table 4.4: Estimated $MISE(\hat{f}_{\mu,\sigma^2}, f_{\mu,\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim U(-\sqrt{3}, \sqrt{3})$, and f_{μ,σ^2} defined by Examples 21-26 of section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different	MISE of f_{μ,σ^2}					
Methods	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0014	0.0002	0.0004	0.0004	0.0139	0.0017
Sample Statistics	0.0553	0.0132	0.4050	0.0284	1.0958	0.1022
EBMLE.XKB	0.0541	0.0115	0.3785	0.0257	0.8234	0.1002
EBMOM.XKB	0.0541	0.0115	0.3756	0.0257	1.0226	0.1006
JS.XKB	0.0552	0.0128	0.3975	0.0274	0.9674	0.0969
SURE.G.XKB	0.0540	0.0112	0.3720	0.0262	1.0932	0.1018
SURE.M.XKB	0.0540	0.0112	0.3748	0.0262	1.0895	0.1015
GL.WMBZ	0.0540	0.0113	0.3594	0.0263	0.7325	0.0961
SURE.M.Double	0.1461	0.0732	0.4838	0.0977	2.0861	0.5094
Histogram	0.0595	0.0162	0.0054	0.0209	0.5777	0.0524

Figure 4.5: Estimated $MISE(\hat{f}_\mu, f_\mu)$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is uniform. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

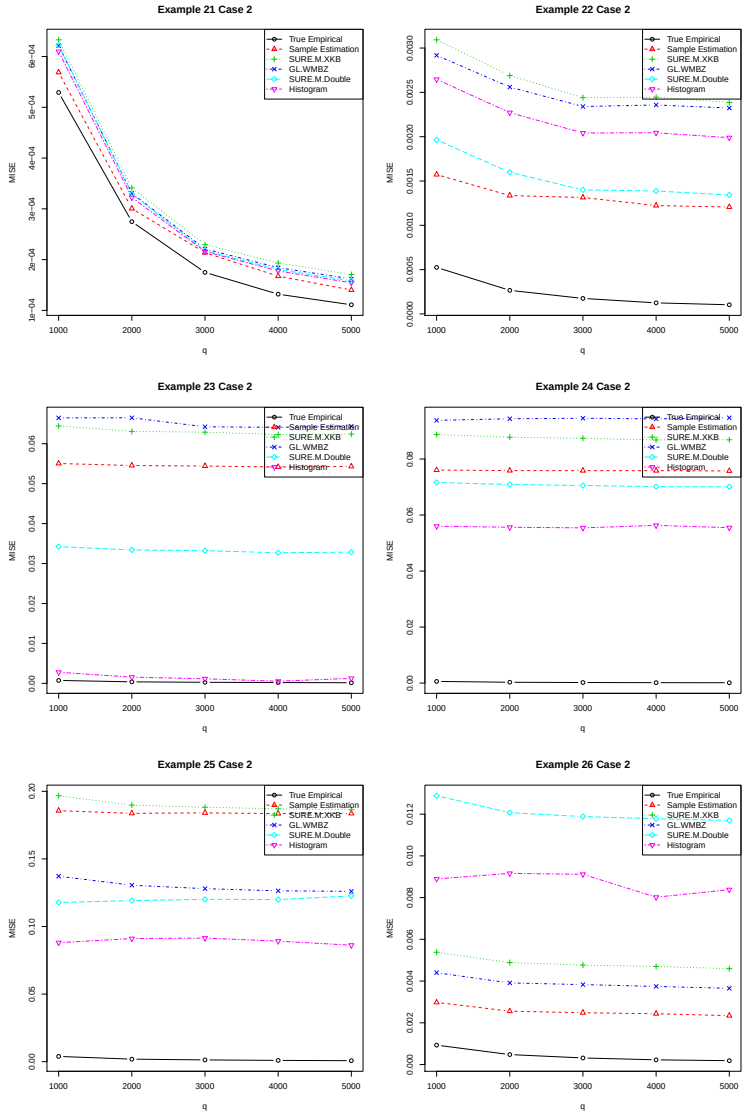


Table 4.5: Estimated $MISE(\hat{f}_\mu, f_\mu)$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim U(-\sqrt{3}, \sqrt{3})$, and f_{μ,σ^2} defined by Examples 21-26 of Section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different	MISE of f_μ					
Methods	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0002	0.0002	0.0003	0.0003	0.0018	0.0004
Sample Statistics	0.0003	0.0013	0.0545	0.0759	0.1841	0.0026
EBMLE.XKB	0.0003	0.0012	0.0363	0.0765	0.1166	0.0040
EBMOM.XKB	0.0003	0.0014	0.0381	0.0781	0.1525	0.0033
JS.XKB	0.0002	0.0004	0.0423	0.0702	0.1221	0.0018
SURE.G.XKB	0.0003	0.0026	0.0502	0.0875	0.1828	0.0026
SURE.M.XKB	0.0003	0.0026	0.0630	0.0875	0.1896	0.0049
GL.WMBZ	0.0003	0.0025	0.0651	0.0944	0.1296	0.0039
SURE.M.Double	0.0003	0.0015	0.0333	0.0707	0.1199	0.0121
Histogram	0.0003	0.0022	0.0015	0.0558	0.0892	0.0087

Figure 4.6: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ vs. dimension q for Examples 21-26 of Section 4.3.1 when f_ϵ is uniform. Dimension size is $q = 1000, 2000, \ldots, 5000$ and number of replications is 100 for each q.

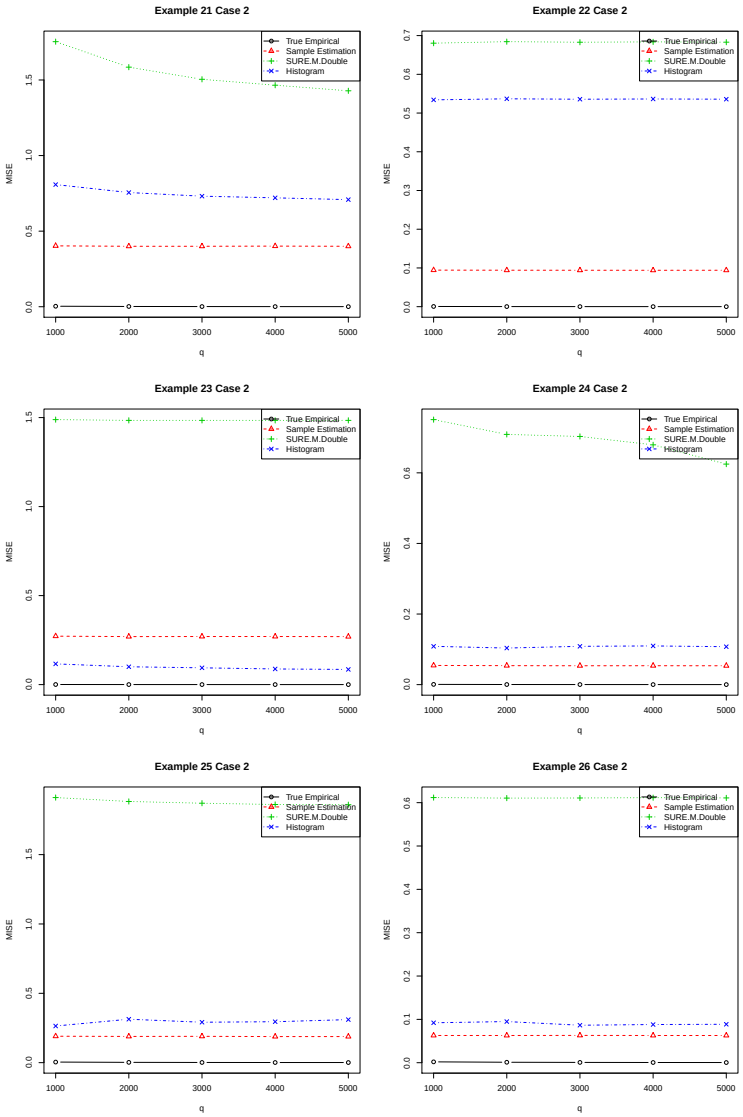


Table 4.6: Estimated $MISE(\hat{f}_{\sigma^2}, f_{\sigma^2})$ averaged over values of q. The data were generated from (1.6) with $n = 4$, $f_\epsilon \sim U(-\sqrt{3}, \sqrt{3})$, and f_{μ,σ^2} defined by Examples 21-26 of Section 4.3.1. At a given q, MISE is estimated by averaging over 100 replications, and then a table value is obtained by averaging over $q = 1000, 2000, \ldots, 5000$.

Different Methods	MISE of f_{σ^2}					
	Example 21	Example 22	Example 23	Example 24	Example 25	Example 26
True empirical	0.0016	0.0002	0.0003	0.0003	0.0019	0.0009
Sample Statistics	0.4007	0.0941	0.2703	0.0537	0.1891	0.0628
SURE.M.Double	1.5473	0.6829	1.4851	0.6934	1.8773	0.6111
Histogram	0.7447	0.5357	0.0968	0.1076	0.1076	0.0881

5. SUMMARY AND CONCLUSIONS

Since Stein's work (Stein [1956]), there has been much progress in using shrinkage estimators of the mean of a high-dimensional normal vector. However, all of the previous work focuses on estimating μ_i and σ_i^2 by minimizing quadratic loss rather than estimating their bivariate density using a mixture that shrinks the sample means in the appropriate direction. We have developed a very general algorithm which does not rely on the belief that all μ_i are of the same magnitude. Our method uses a mixture of normal-inverse gamma densities to estimate the bivariate density, f_{μ,σ^2}. This method effectively clusters sample means into different groups and then shrinks an individual mean towards its corresponding group mean. Our algorithm outperforms SURE methods in terms of squared error loss when μ_i and σ_i^2 are dependent, and outperforms group linear algorithms in terms of squared error loss when μ_i and σ_i^2 are independent. When μ_i has a multimodal distribution or when σ_i^2 is unknown, our method based on mixtures of normal-inverse gamma distributions performed better than all the other methods with which it was compared.

Also, our approach allows us to estimate the joint density of (μ_i, σ_i^2), a problem which seems not to have been previously addressed. Our algorithm outperforms SURE methods when we plug estimated μ_i and σ_i^2 into kernel density estimators. In some extreme cases where our $N\Gamma^{-1}$ mixture does not perform well, $N\Gamma^{-1}$ KDE always outperforms the other plug-in estimators.

When f_ϵ is not normal then we may not use $N\Gamma^{-1}$ mixture to estimate the joint density of (μ, σ^2). We develop an algorithm based on a bivariate histogram which can estimate the density f_{μ,σ^2} for any known f_ϵ. This algorithm outperforms other methods which are based on "plug-in KDE" when the true density of (μ, σ^2) is not smooth. We find that when f_{μ,σ^2} is smooth, "plug-in KDE" is better estimates of the density f_{μ,σ^2} compare to bivariate histogram as the bivariate histogram is not smooth. However, bivariate histogram outperforms "plug-in histogram estimates" in most situations.

REFERENCES

Alvin J Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, pages 642–645, 1970.

Rudi Beran and P Warwick Millar. Minimum distance estimation in random coefficient regression models. *The Annals of Statistics*, pages 1976–1992, 1994.

James O Berger. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *The Annals of Statistics*, pages 223–226, 1976.

James O Berger and William E Strawderman. Choice of hierarchical priors: admissibility in estimation of normal means. *The Annals of Statistics*, pages 931–951, 1996.

Lawrence D Brown. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics*, 42(3):855–903, 1971.

Lawrence D Brown. In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, pages 113–152, 2008.

Lawrence D Brown and Eitan Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704, 2009.

Raymond J Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

Raymond J Carroll and Peter Hall. Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):31–46, 2004.

Aurore Delaigle and Peter Hall. Methodology for non-parametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):231–252, 2016.

Aurore Delaigle, Peter Hall, and Alexander Meister. On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685, 2008.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitorsâĂŤan empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Jianqing Fan. Global behavior of deconvolution kernel estimates. *Statistica Sinica*, pages 541–551, 1991.

Jianqing Fan. Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169, 1992.

Thomas S Ferguson. Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24(1983):287–302, 1983.

Peter Hall and Yanyuan Ma. Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):429–446, 2007.

Jeffrey D Hart and Isabel Cañette. Nonparametric estimation of distributions in random effects models. *Journal of Computational and Graphical Statistics*, 20(2):461–478, 2011.

Joel L Horowitz and Marianthi Markatou. Semiparametric estimation of regression models for panel data. *The Review of Economic Studies*, 63(1):145–168, 1996.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, pages 361–379, 1961.

Wenhua Jiang, Cun-Hui Zhang, et al. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

Bing-Yi Jing, Zhouping Li, Guangming Pan, and Wang Zhou. On sure-type double shrinkage estimation. *Journal of the American Statistical Association*, 111(516):1696–1704, 2016.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Tong Li and Quang Vuong. Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65(2):139–165, 1998.

Xihong Lin and Raymond J Carroll. Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):69–88, 2006.

Bruce G Lindsay et al. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, 11(1):86–94, 1983.

Julie McIntyre and Leonard A Stefanski. Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics*, 63(1):81–99, 2011.

Omkar Muralidharan. An empirical bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, pages 422–438, 2010.

Olav Reiersøl. Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society*, pages 375–389, 1950.

Abhra Sarkar, Bani K Mallick, John Staudenmayer, Debdeep Pati, and Raymond J Carroll. Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics*, 23 (4):1101–1125, 2014.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

John Staudenmayer, David Ruppert, and John P Buonaccorsi. Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, 103(482):726–736, 2008.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal

distribution. Technical report, STANFORD UNIVERSITY STANFORD United States, 1956.

Zhiqiang Tan et al. Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli*, 21(1):574–603, 2015.

Henry Teicher. On the mixture of distributions. *The Annals of Mathematical Statistics*, pages 55–73, 1960.

Asaf Weinstein, Zhuang Ma, Lawrence D Brown, and Cun-Hui Zhang. Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*, pages 1–13, 2018.

Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, pages 75–88, 1957.

Xianchao Xie, SC Kou, and Lawrence D Brown. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479, 2012.

Xianyang Zhang and Anirban Bhattacharya. Empirical bayes, sure and sparse normal mean models. *arXiv preprint arXiv:1702.05195*, 2017.

APPENDIX A

COMPUTATIONS FOR UNIFORM SCALED ERROR

In Section 4.1.2.2 we discussed how to compute the integral J_{ir} for all cases. There are a total of 6 different cases for how two lines can intersect a box in the plane of (m, s).

Case 1: Rectangle entirely outside the region $m < X_{i(1)} + sc$ and $m > X_{i(n)} - sc$.

Case 2: Rectangle entirely inside the region $m < X_{i(1)} + sc$ and $m > X_{i(n)} - sc$.

Case 3: Rectangle intersects $m = X_{i(n)} - sc$ but not $m = X_{i(1)} + sc$. (4 subcases)

Case 4: Rectangle intersects $m = X_{i(1)} + sc$ but not $m = X_{i(n)} - sc$. (4 subcases)

Case 5: Rectangle intersects both lines and intersection is inside rectangle. (4 subcases)

Case 6: Rectangle intersects both lines and intersection is outside rectangle. (4 subcases)

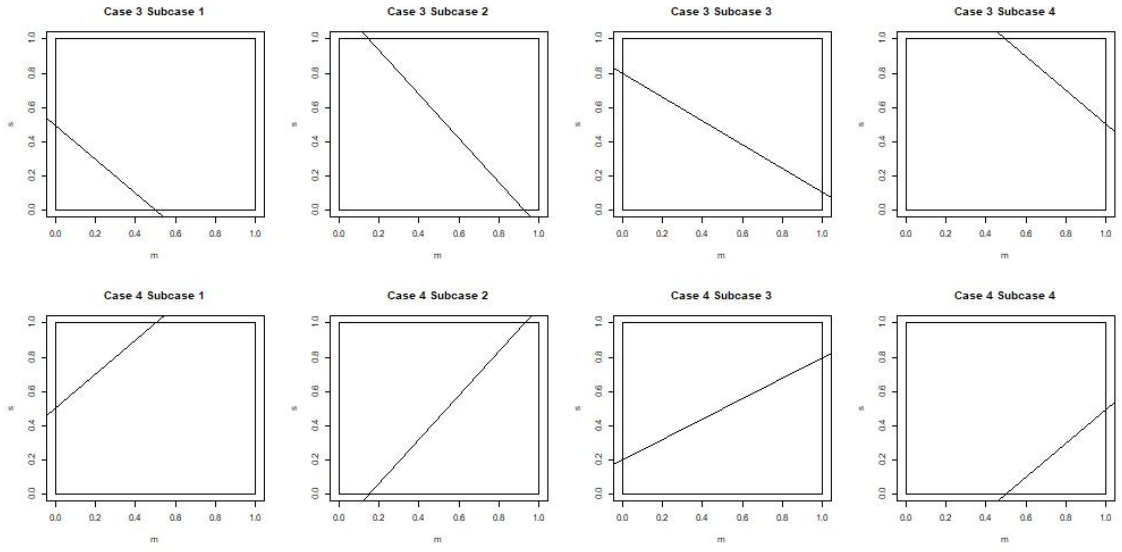


Figure A.1: 4 different subcases of Case 2 and Case 3 when estimating f_{μ,σ^2} using a histogram and f_ϵ is uniform as discussed in Section 4.1.2.2.

Let $T_{ir} = 2^{n-1}c^n J_{ir}$ and $d = n - 1$

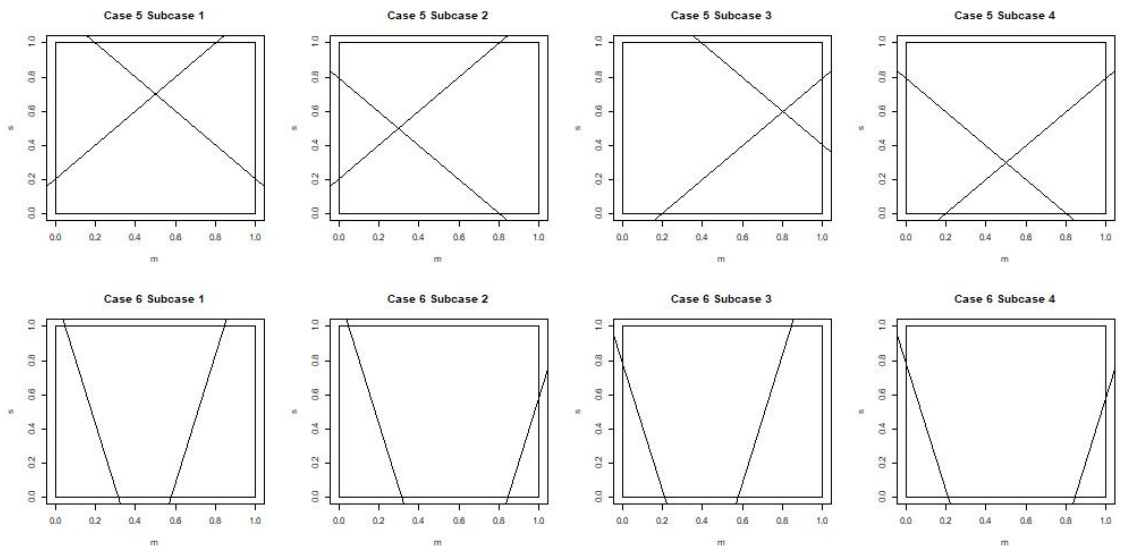


Figure A.2: 8 different subcases of Case 5 when estimating f_{μ,σ^2} using a histogram and f_ϵ is uniform as discussed in Section 4.1.2.2.

$$T_{ir} = \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{s-1}}^{a_s} (s)^{-d} I_{(X_{i(n)}-sc, X_{i(1)}+sc)}(m) dm ds$$

Case 1:

When $a_j < X_{i(n)} - c\sqrt{b_t}$ or $a_{j-1} > X_{i(1)} + c\sqrt{b_t}$ or $\sqrt{b_t} < (X_{i(n)} - X_{i(1)})/(2c)$ then

$$T_{ir} = 0$$

Case 2:

When $a_{j-1} > X_{i(n)} - c\sqrt{b_{t-1}}$ and $a_j < X_{i(1)} + c\sqrt{b_{t-1}}$ then

$$T_{ir} = \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{j-1}}^{a_j} s^{-d} dm ds = (a_j - a_{j-1})/(d-1)[(\sqrt{b_{t-1}})^{-d+1} - (\sqrt{b_t})^{-d+1}]$$

Case 3:

When $a_{j-1} < X_{i(n)} - c\sqrt{b_{t-1}}$ and $a_j > X_{i(n)} - c\sqrt{b_t}$ and $(a_j < (X_{i(1)} + X_{i(n)})/2$ or $a_j < X_{i(1)} + c\sqrt{b_{t-1}})$ then

Subcase 1

$$T_{ir} = \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{\max(a_{j-1}, X_{i(n)}-sc)}^{a_j} s^{-d} dmds$$
$$= (X_{i(n)} - a_j)/(d-1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] -$$
$$c/(d-2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] +$$
$$(a_j - a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}]$$

Subcase 2

$$T_{ir} = \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{X_{i(n)}-sc}^{a_j} s^{-d} dmds = (X_{i(n)} - a_j)/(d-1)[(\sqrt{b_t})^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] -$$
$$c/(d-2)[(\sqrt{b_t})^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}]$$

Subcase 3

$$T_{ir} = \int_{a_{j-1}}^{a_j} \int_{(X_{i(n)}-m)/c}^{\sqrt{b_t}} s^{-d} dsdm$$
$$= (a_j - a_{j-1})/(d-1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} - (\sqrt{b_t})^{-d+1}] +$$
$$(X_{i(n)} - a_j)/(d-1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} -$$
$$((X_{i(n)} - a_j)/c)^{-d+1})] + c/(d-2)[((X_{i(n)} - a_j)/c)^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}]$$

Subcase 4

$$T_{ir} = \int_{(X_{i(n)}-sc}^{\sqrt{b_t}} \int_{X_{i(n)}-sc}^{a_j} s^{-d} dmds$$
$$= (X_{i(n)} - a_j)/(d-1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_j)/c)^{-d+1}] -$$
$$c/(d-2)[((X_{i(n)} - a_j)/c)^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}]$$

Case 4:

When $a_{j-1} < X_{i(1)} + b_k c$ and $a_j > X_{i(1)} + c\sqrt{b_{t-1}}$ and $(a_{j-1} > (X_{i(1)} + X_{i(n)})/2$ or $a_{j-1} > X_{i(n)} - c\sqrt{b_{t-1}})$ then

Subcase 1

$$\begin{aligned} T_{ir} &= \int_{(a_{j-1}-X_{i(1)})/c}^{\sqrt{b_t}} \int_{a_{j-1}}^{X_{i(1)}+sc} s^{-d} dm ds \\ &= (a_{j-1} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1)} - ((a_{j-1} - X_{i(1)})/c)^{-d+1}] - \\ & c/(d-2)[(\sqrt{b_t})^{-d+2} - ((a_{j-1} - X_{i(1)})/c)^{-d+2}] \end{aligned}$$

Subcase 2

$$\begin{aligned} T_{ir} &= \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{j-1}}^{X_{i(1)}+sc} s^{-d} dm ds \\ &= (a_{j-1} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] - \\ & c/(d-2)[(\sqrt{b_t})^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \end{aligned}$$

Subcase 3

$$\begin{aligned} T_{ir} &= \int_{a_{j-1}}^{a_j} \int_{(m-X_{i(1)})/c}^{\sqrt{b_t}} s^{-d} ds dm \\ &= (a_{j-1} - X_{i(1)})/(d-1)[((a_j - X_{i(1)})/c)^{-d+1} - ((a_{j-1} - X_{i(1)})/c)^{-d+1}] - \\ & c/(d-2)[((a_j - X_{i(1)})/c)^{-d+2} - ((a_{j-1} - X_{i(1)})/c)^{-d+2}] + \\ & (a_j - a_{j-1})/(d-1)[((a_j - X_{i(1)})/c)^{-d+1} - (\sqrt{b_t})^{-d+1}] \end{aligned}$$

Subcase 4

$$T_{ir} = \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{a_{j-1}}^{\min(X_{i(1)}+sc, a_j)} s^{-d} dmds$$

$$= \int_{\sqrt{b_{t-1}}}^{(a_j - X_{i(1)})/c} \int_{a_{j-1}}^{X_{i(1)}+sc} s^{-d} dmds + \int_{(a_j - X_{i(1)})/c}^{\sqrt{b_t}} \int_{a_{j-1}}^{a_j} s^{-d} dmds$$

$$= (a_j - a_{j-1})/(d-1)[((a_j - X_{i(1)})/c)^{-d+1} - (\sqrt{b_t})^{-d+1}] +$$

$$(X_{i(1)} - a_{j-1})/(d-1)[(\sqrt{b_{t-1}})^{-d+1} - ((a_j - X_{i(1)})/c)^{-d+1}] -$$

$$c/(d-2)[((a_j - X_{i(1)})/c)^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}]$$

Case 5:

When $(X_{i(n)} - X_{i(1)})/(2c) > (\sqrt{b_{t-1}})$ and $(X_{i(n)} - X_{i(1)})/(2c) < (\sqrt{b_t})$ and

$(X_{i(n)} + X_{i(1)})/2 > a_{j-1}$ and $(X_{i(n)} + X_{i(1)})/2 < a_j$ then

Subcase 1

$$T_{ir} = \int_{(X_{i(n)} - X_{i(1)})/(2c)}^{\sqrt{b_t}} \int_{X_{i(n)}-sc}^{X_{i(1)}+sc} s^{-d} dmds$$

$$= (X_{i(n)} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+1}] -$$

$$2c/(d-2)[(\sqrt{b_t})^{-d+2} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+2}]$$

Subcase 2

$$T_{ir} = \int_{(X_{i(n)} - X_{i(1)})/(2c)}^{\sqrt{b_t}} \int_{X_{i(n)}-sc}^{X_{i(1)}+sc} s^{-d} dmds + \int_{(X_{i(n)} - a_{j-1})/c}^{\sqrt{b_t}} \int_{a_{j-1}}^{X_{i(1)}+sc} s^{-d} dmds$$

$$= (X_{i(n)} - X_{i(1)})/(d-1)[(X_{i(n)} - a_{j-1})/c)^{-d+1} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+1}] +$$

$$2c/(-d+2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+2}] +$$

$$(X_{i(1)} - a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}] -$$

$$c/(d-2)[(\sqrt{b_t})^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}]$$

Subcase 3

$$T_{ir} = \int_{(X_{i(n)}-X_{i(1)})/(2c)}^{(a_{j-1}-X_{i(1)})/c} \int_{X_{i(n)}-sc}^{X_{i(1)}+sc} s^{-d} dmds + \int_{(X_{i(n)}-a_{j-1})/c}^{\sqrt{b_t}} \int_{X_{i(n)}-sc}^{a_j} s^{-d} dmds$$

$$= (X_{i(n)} - X_{i(1)})/(d-1)[(X_{i(n)} - a_{j-1})/c)^{-d+1} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+1}]+$$

$$2c/(-d+2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+2}]+$$

$$(X_{i(1)} - a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}]-$$

$$c/(d-2)[(\sqrt{b_t})^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}]$$

Subcase 4

$$T_{ir} = \int_{a_{j-1}}^{a_j} \int_{\max((X_{i(n)}-m)/c,(m-X_{i(1)})/c)}^{\sqrt{b_t}} s^{-d} dsdm$$

$$= I((a_{j-1} + a_j) < (X_{i(1)} + X_{i(n)}))$$

$$\Big\{(X_{i(n)} - X_{i(1)})/(d-1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+1}]+$$

$$2c/(-d+2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+2}]+$$

$$(X_{i(1)} - a_{j-1})/(-d+1)[((a_j - X_{i(1)})/c)^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}]+$$

$$c/(-d+2)[((a_j - X_{i(1)})/c)^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}]+$$

$$(a_j - a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((a_j - X_{i(1)})/c)^{-d+1}] + \Big\}$$

$$I((a_{j-1} + a_j) > (X_{i(1)} + X_{i(n)}))$$

$$\Big\{(X_{i(n)} - X_{i(1)})/(d-1)[((a_j - X_{i(1)})/c)^{-d+1} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+1}]+$$

$$2c/(-d+2)[((a_j - X_{i(1)})/c)^{-d+2} - ((X_{i(n)} - X_{i(1)})/(2c))^{-d+2}]+$$

$$(X_{i(1)} - aj - 1)/(-d+1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} - ((a_j - X_{i(1)})/c)^{-d+1}]+$$

$$c/(-d+2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - ((a_j - X_{i(1)})/c)^{-d+2}]+$$

$$(a_j - a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}]\Big\}$$

Case 6:

When $(\sqrt{b_{t-1}}) > (X_{i(n)} - X_{i(1)})/(2c)$ and $a_{j-1} < X_{i(n)} - c\sqrt{b_{t-1}}$ and $a_j > X_{i(n)} - c\sqrt{b_t}$ and $a_{j-1} < X_{i(1)} + c\sqrt{b_t}$ and $a_j > X_{i(1)} + c\sqrt{b_{t-1}}$ then

Subcase 1

$$\begin{aligned} T_{ir} &= \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{X_{i(n)}-s/c}^{X_{i(1)}+s/c} s^{-d} dmds \\ &= (X_{i(n)} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] - \\ &2c/(d-2)[(\sqrt{b_t})^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \end{aligned}$$

Subcase 2

$$\begin{aligned} T_{ir} &= \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} \int_{X_{i(n)}-s/c}^{X_{i(1)}+s/c} s^{-d} dmds \\ &= (X_{i(n)} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] - \\ &2c/(d-2)[(\sqrt{b_t})^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \end{aligned}$$

Subcase 3

$$\begin{aligned} T_{ir} &= \int_{\max(a_{j-1}, X_{i(n)}-sc)}^{X_{i(1)}+sc} \int_{\sqrt{b_{t-1}}}^{\sqrt{b_t}} s^{-d} dsdm \\ &= (X_{i(n)} - X_{i(1)})/(d-1)[((X_{i(n)} - a_{j-1})/c)^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] \\ &+ 2c/(-d+2)[((X_{i(n)} - a_{j-1})/c)^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \\ &+ (a_{j-1} - X_{i(1)})/(d-1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)} - a_{j-1})/c)^{-d+1}] \\ &+ c/(-d+2)[(\sqrt{b_t})^{-d+2} - ((X_{i(n)} - a_{j-1})/c)^{-d+2}] \end{aligned}$$

Subcase 4

$$
\begin{aligned}
T_{ir} &= \int_{a_{j-1}}^{a_j} \int_{\max((X_{i(n)}-m)/c,(m-X_{i(1)})/c,(\sqrt{b_{t-1}}))}^{\sqrt{b_t}} s^{-d} ds dm \\
&= I((a_{j-1}+a_j) < (X_{i(1)}+X_{i(n)})) \\
&\Big\{ (X_{i(n)}-X_{i(1)})/(d-1)[((X_{i(n)}-a_{j-1})/c)^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] \\
&+ 2c/(-d+2)[((X_{i(n)}-a_{j-1})/c)^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \\
&+ (a_{j-1}-X_{i(1)})/(d-1)[((a_j-X_{i(1)})/c)^{-d+1} - ((X_{i(n)}-a_{j-1})/c)^{-d+1}] \\
&+ c/(-d+2)[((a_j-X_{i(1)})/c)^{-d+2} - ((X_{i(n)}-a_{j-1})/c)^{-d+2}] \\
&+ (a_j-a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((a_j-X_{i(1)})/c)^{-d+1}] \Big\} \\
&+ I((a_{j-1}+a_j) > (X_{i(1)}+X_{i(n)})) \\
&\Big\{ (X_{i(n)}-X_{i(1)})/(d-1)[((a_j-X_{i(1)})/c)^{-d+1} - (\sqrt{b_{t-1}})^{-d+1}] \\
&+ 2c/(-d+2)[((a_j-X_{i(1)})/c)^{-d+2} - (\sqrt{b_{t-1}})^{-d+2}] \\
&+ (X_{i(n)}-a_j)/(d-1)[((X_{i(n)}-a_{j-1})/c)^{-d+1} - ((a_j-X_{i(1)})/c)^{-d+1}] \\
&+ c/(-d+2)[((X_{i(n)}-a_{j-1})/c)^{-d+2} - ((a_j-X_{i(1)})/c)^{-d+2}] \\
&+ (a_j-a_{j-1})/(-d+1)[(\sqrt{b_t})^{-d+1} - ((X_{i(n)}-a_{j-1})/c)^{-d+1}] \Big\}
\end{aligned}
$$